MASTER IN INNOVATION AND RESEARCH IN INFORMATICS
Specialty: Computer Networks and Distributed Systems

Multivariate statistical analysis of urban environments applied to New York City

(A contextual analysis of urban semantics)

Presented by: Cedric K. Bhihe
<cedric.bhihe@gmail.com>

Defended on: January 30, 2019

Advisor: Prof. Jorge García Vidal
Dept of Computer Architecture
Co-advisor: Prof. José Mª Barceló Ordinas
Dept of Computer Architecture

FACULTAT D'INFORMÀTICA DE BARCELONA (FIB)
UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC)
BarcelonaTech

Abstract

This work is but a modest contribution to the broad research domain of SmartCities and mobility networking. It is primarily aimed at city officials and other decision-makers, in an effort to help them assign budgetary and human resources based on a more reliable multidimensional data representation in smart urban environments.

We attempt to show how sensing weak signals, namely service request calls (SRCs) logged by the New York City (NYC) municipality, enriched with crime report calls (CRCs) from city dwellers, may help in:

(a) appraising the "mood" of city dwellers subject to a changing urban context, and
(b) characterizing the space and time distribution of urban pathologies (as perceived/reported by callers)

We primarily rely on multivariate statistical analysis of low latency urban data (our so-called weak signals) and on the comparative analysis of successive time windows to understand crime and its statistical correlations to the perception of other urban factors. We delineate the limitations of the analytical framework used to visualize complex, multidimensional data in urban settings. As a conclusion to this exploratory work, we propose to extend it with a new structured research effort to build a more satisfactory visualization framework.

The composite data set used in this work comprises data originating from different digital sources. We produced a mostly automated ETL pipeline capable of processing complex, composite data almost unattended. It yields geo-localized information organized as a count table by ZIP codes (rows) and categorical variables' modalities (columns) in the 5 historical boroughs which make up New York City. Data mining revealed that data is all too often incomplete, sometimes wrong, or statistically unreliable. Statisticians are keenly aware that skewing data, or introducing bias(es) at any processing stage is a deadly *caveat emptor*. We therefore exerted caution not to introduce data biases or, when inevitable, to characterize it properly.

First, linear dimensional reduction methods such as PCA, CA and even MCA helped us garner information about the hidden structure of our data sets. Based on the determination of directions of maximum variance, coupled with feature selection and extraction, latent feature analysis was proposed at various stages of the work. We confirmed a well-known result, namely that low frequency cells have a dramatic impact on visualized results. This led us to rid our data set from such spurious effects whenever possible.

Second, we deployed generic tools of clustering, and identified a number of cluster classes, which varied according to the data sets' time-windows. Clustering results do not coincide topologically with NYC's five boroughs, but rather with particular traits of the local geography and of the residents' socio-economic makeup inside those boroughs. Factors and intensity levels of explicit factors' modalities, instrumental in the statistical construction of cluster classes, were elucidated. We showed how dramatic the effect of the addition of crime rate was, as it shifted the focus away from Manhattan and onto other boroughs, in terms of variance explanatory power.

Third we produce change maps in the form of heat-maps based on the Mahalanobis distance between observed items from one data set time window to the next.

Last we propose an extension to this exploratory work, in the form of a research program. The program focuses on non-linear dimensionality reduction (NLDR), and manifold learning to capture more of the hidden structure at close and medium range, rather than at long range as is the case with conventional linear methods such as PCA, CA and MCA. In this proposal visualization becomes a means to gain further insight in complex, multi-dimensional urban data by incorporating different measures of proximity in space and time. We envision further possible extensions, to include rich text semantic analysis applied to urban events and points of interests (POIs).

■■■

Acknowledgments

I am indebted to my family for its extraordinary support of a very personal decision, which consisted in returning to school after a good twenty-five years of non-academic professional endeavors in more countries and on more continents than I care to mention here. Ana, my accomplice in life, deserves a very special mention in that area, for her unwavering support, her love and her renewed patience throughout the two year MIRI program at UPC.

Having been accepted at UPC's MIRI program as an "older student" in the first place, means that Prof. José Mª Barceló Ordinas, who heads the Computer Network and Distributed Computing (CNDS) area's selection process within MIRI, saw some merit in my application to the Master degree program. In retrospect his decision made all the rest possible. I am sincerely grateful for that.

Both Prof. Jorge García Vidal and Prof. José Mª Barceló Ordinas are dedicated researchers and lecturers of the Department of Computer Architecture (DAC). For the graduate level subject matters they teach (respectively statistical analysis applied to networks and game theory), they spare no effort in showing students what is essential in the pursuit of new knowledge: namely to develop a good grasp of the basic underlying mechanisms and governing ideas subtending theoretical or experimental approaches. As my two co-advisors for this work, they also proved to be an invaluable source of advice, with a focus on pertinence and applicability which trumped all my expectations. Each brings "connecting the dots" to a new level and humor to where it should always be. For all those reasons working *with* them and *under* their guidance has been a pleasure and a privilege. I am extremely thankful to both for that. At time of writing we co-authored a research proposal along with others to extend this work along the main lines of Section 5 in this report.

Unavoidably learning involves a learner's mind and heart, some learning material, i.e. organized information destined to become knowledge, as well as a process of transmission, with as its origin an author and/or a teacher. In that latter area I would like to mention Prof. Tomàs Aluja Banet, who taught me pretty much all I dare think I understand about multivariate analysis, Profs. Argimiro Arratia and Ramon Ferrer i Cancho for their stimulating lectures on graphs of complex and social networks, Prof. Marta Arias (as an unexpected substitute for the temporarily indisposed originally programmed lecturer) for her broad introduction to machine learning, Prof. Jordi Domingo Pascual for his communicative enthusiasm about and insights in computer networks and the future of the Internet, Profs. Leandro Navarro and Jordi Torres for helping and cheering along as I made my first baby-steps in the area of decentralized systems and cloud computing, and, last but not least, Prof. Llorenç Cerdà Alabern for his fascinating lectures on stochastic network modelling and how one may (or may not) usefully approach non-deterministic processes.

Beginnings are difficult at times, new beginnings sometimes more so. I cannot say that I loved every moment spent studying within MIRI; time was mostly scarce and the volume of new information was staggering. However smoothed and averaged over a two month long sliding time-window, the assertion definitely holds. During some of the early uneasy times I was very fortunate to have Lucas Robin, Guillaume Lebrun and Toni Pohl, as fellow students and soon as friends, coincidentally all enrolled in the MIRI program as Erasmus students. Their help in explaining certain arcane aspects of either object-oriented, functional or distributed programming was invaluable and I want to express my gratitude to them for it. Lucas is now a patent expert specialized in AI, Guillaume leads complex projects in a consulting firm and Toni is a large scale computer research scientist at IBM in Germany.

I was not so rusty after all, even after twenty-five plus years away from academia. However, roughly thirty months ago the various domains of Computer Science were *terra incognita* to me, as I came from Physics and Engineering Mechanics. My notions of Computer Science were as naive as could be for a physicist born in 1963, perhaps even laughable. In any case that fact added a spicy element of surprise and discovery to a 2 year journey filled with study. It spurred me to dig deeper whenever the opportunity arose. I was always encouraged in that by a good friend and renowned machine learning specialist, Prof. Ricard Gavaldà Mestre, whom at last I wish to thank dearly here.

■■■ Barcelona, Dec. 2018

Table of Contents

Foreword

A few years ago, Ben Wellington published an article[i] about mapping New York City's noisiest neighborhoods, soon followed by another one producing results on the hidden circumstances behind New York City's permanent traffic gridlock[ii]. Those two articles, published in the New Yorker, were meant for a wide (although somewhat upscale) readership. They also revealed that the author had used analytical and statistical methods based on a rich data base. That database is *NYC Open Data*[iii], a trove of information geographically and temporally more precise than census tract scale data as made publicly available by the US Federal Government. We tapped it. This report describes to what ends, to what extent and how.

Dense urban areas are usually complex environments, characterized by a large, heterogeneous set of co-varying quantities which put together (in time and in space) constitute the *urban semantics*. As such the fabric of cities is difficult to understand by both businesses and city government officials. In the face of sometimes conflicting priorities and difficult to grasp multi-dimensional issues, businesses and municipalities often rely on empirical data. That data (however incomplete or biased) becomes the basis for intuitive, non-explicit and unverified correlations, the which may lead to flawed decisions and later to corrective actions. Data analysts might well be able to do better on both counts. This work suggests how.

Today's paradox is that the out pour of available smart-city data is often too much, too heterogeneous or too complex to be usefully tapped and visualized by either the governments of the very smart cities at the origin of the data or organizations with a vested interest in exploiting it… and so, urban semantics remain an idiom difficult to understand. The data in question may be dynamic or static. The former may include calls to 911, calls to 311, car traffic, weather, accidents, social media comments, etc. They are signals whose update or accrual frequency is large and are therefore often dubbed "low latency" or "weak" signals". The order of magnitude of their update period may range from 1 second to under 50 hours. By contrast, the latter (static data) consists of "high latency" urban data deriving from census data, income tax, unemployment, political leaning or from a city's urban landscape, such as the presence of points of interest (POIs). We dub them strong signals. Their update period may range from days to years.

Weak signals (low latency data) are harvested continuously by different agencies, municipal entities and private networks. For data to be accessible to us, it must be stored digitally in such a way that its posterior analysis is possible. Ultimately at stake is for city officials and businesses alike to better grasp urban semantics, their evolution and predictability. Decision-makers have common objectives: to make better decisions, to build better strategies and better policy, on which to base an optimal allocation of resources both in time and space.

The number of connected municipalities across the world, likely interested in better allocating their resources, is understandably large. The continued influx of people in cities make predictive management a sensitive must-have, once tools become available. The increasing size of modern conurbations has direct consequences in terms of the emerging complexity of its semantics. That makes the ability to manipulate big data in an automated way and to visualize its hidden structure attractive. Beneficiaries include both business people seeking to maximize their ROI and city officials seeking to maximize both the well-being of inhabitants under their responsibility… and their chances of staying in office.

Cities becoming bigger and attracting more people year after year constitutes a trend, which has consolidated over the past one hundred years. The need for optimal resource allocation and complex business decision-making should continue to assert itself, reinforced by increased environmental stress on dense cities due to global warming. It may be counterbalanced, at least in part, by the long term possibility of large swaths of urban populations leaving their urban environments. This long term scenario (30 years in the future at least) finds its roots in soaring living costs in large cities. It does not constitute an actual threat to this project, and should therefore not detract from its purported timeliness and usefulness.

The key issue, as often perceived by the potential beneficiaries of this exploratory work, is how to best interpret urban semantics. After extracting available urban data, we tackle the issue by automating data transformation to the extent possible, before conducting both linear and non-linear statistical and machine learning. We first used conventional methods largely based on PCA, CA, MCA and k-means consolidated hierarchical clustering. Last we propose an heterogeneous urban data visualization framework, so that to reveal the hidden structure of urban data in a way accessible to citizens and decision makers. In this framework observation proximity, both in time and space, is the basis for spatial as well as associative inferences, in a way conducive to the further elaboration of what-if scenarios.

---

i https://www.newyorker.com/tech/elements/mapping-new-york-noise-complaints

ii https://www.newyorker.com/tech/elements/uber-isnt-causing-new-york-citys-traffic-slowdown

iii https://opendata.cityofnewyork.us/

1. Introduction

Since 2010, between 2,500 and 15,000 daily calls to 311 are recorded in New York City, NY (NYC). Those service request calls (SRCs) are logged with a slew of attributes (more than 50 fields are available per call), on the location of the reported incident, its nature (e.g. noise, public housing conditions, street potholes, stray animals, rodent sighting, ailing trees, barking dogs, unsanitary food establishments, uncivil behavior, parking violations, etc.). SRCs' attributes include time and date , as well as geo-location of the incident, reasons and object of the call. They are available[iv] either freely on the Internet by courtesy of the municipal government of NYC, or otherwise covered by the US FOIA[v].

Simultaneously the NYPD, New York's Police Department, registers over 1000 daily felonies, misdemeanors and violations[vi]. This affords the curious analyst a rich overview on the type of issues being reported, their location, and frequency. It is also an invitation to scrutinize possible correlations between the statistics of geo-located 311 SRCs and other factors such as population density, type of criminality, median income and IRS declared jobless benefits in income tax returns. We will restrict our geographical reach to ZIP code areas of neighborhoods in the 5 boroughs of NYC per Figure 1: Manhattan, Brooklyn, Queens, the Bronx, and Staten Island. All other ZIP codes are excluded.

***Figure 1**: NYC's five historical boroughs (source: Wikipedia)*

Dealing with city areas tabulated by ZIP[vii] code is in general less precise than doing so with census tracts as ZIP code's topographical areas are subject to change with time. Census tracts also tend to represent a finer topological mesh than ZIP codes areas as there are approximately 50% more census tracts in the US as there are ZIP codes, a ratio with roughly holds for NYC. We further discuss and justify the choice of the ZIP code area for this survey in a short discussion at the very beginning of Section 3.

In the end, curiosity is what really subtends every human endeavor. More specifically in our case, the motivation to embark on this study was:

(i) to evaluate how much insight can be gained from realistic multidimensional data using classical multi-variate analysis (MVA) exploratory tools,

(ii) to explore, outline and perhaps even implement a more robust visualization framework, capable of accounting for complex multidimensional data in a way consistent with very heterogeneous data. What passes as "heterogeneous" is defined later. We initially thought of a relatively recent Machine Learning technique, *t-SNE*, in a effort to capture our urban data's hidden structure.

In a first part (Sections 2 to 4), we present results based on Correspondence Analysis (CA), Principal Component Analysis (PCA), Clustering and Multiple Correspondence Analysis (MCA) to conduct data exploration and feature extraction. Whenever suitable an effort is made to also offer a critical discussion of obtained results. In a second part (Sections 5 and 6), we propose an implementation of *t-SNE* and/or *UMAP*, two non-linear dimensional reduction techniques, to reveal some more of the hidden structure in urban data.

A less theoretically minded question is ultimately to reveal evolution patterns in the urban fabric of NYC. Our objective is to try to extract predictor-variables on the scale of a ZIP code area. Possible applications are many:

- to predict crime, or at least to establish strong correlations between crime and other urban events and signals.

---

iv *https://data.cityofnewyork.us/Social-Services/311-Service-Requests/fvrb-kbbt*

v *The Freedom Of Information Act is a companion to the US Privacy Act of 1974 (5 U.S.C. 552a). Under the FOIA, anyone residing legally in the USA can make a request for a Federal Agency record.*

vi *https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243*

vii *ZIP or "Zone Improvement Plan" is a territorial mapping used by the US Postal Service (USPS) for snail mail delivery since 1963.*

- to link complaints about urban nuisance to certain neighborhoods and illustrate those neighborhoods in terms of social-economical categories,

- to produce the basis reference model to help decide where to locate what business for maximum attractiveness to customers and return on investment for investors,

- to optimize resources to better manage dense urban areas.

Although we provide a table of contents (p. 4 of this document), a brief description of how this report is organized follows:

- Section 2, we present the protracted process of extracting data from various databases. This included cleaning it (in particular in terms of missing values) and modifying it from a time record format to a location oriented frequency table. Data cleaning, while not intrinsically or conceptually difficult, is a task laden with traps. It occupied over 170 hours of our time. This section sheds light on why and how. It can be skipped and the reader may go directly to the analysis of Section 3.

- Section 3, encompasses different aspects of the multivariate data analysis including CA and PCA on NYC311 SRCs, Clustering and MCA on 2 categorical variables (SRCs and CRCs), plus 1 (illustrative) supplementary variable and 2 quantitative variables. The initial analysis is presented in details for the April 2014 data set and the two already mentioned categorical variables accounting for 16 modalities. Results for the April 2010 and April 2018 data set are included as Appendices.

- Section 4, offers a conclusion on obtained MVA and clustering results and suggests new directions to pursue this multivariate analysis.

- Finally, in Sections 5 to 6, we cover the proposed implementation of two manifold learning technique, of the non-parametric kind (e.g. *t-SNE*), and of the parametric kind (e.g. *UMAP*) in order to tackle visualization issues specific to high dimensional, heterogeneous urban data.

Due to external constraints imposed on this work, results presented in Sections 1 to 4 of this report were obtained exclusively by relying on custom R scripts. Notwithstanding external constraints, we cannot but warmly advise interested coders, not to code with R during the data cleaning phase. R is quirky at times, and has either scant or too much incomplete documentation. Being FOSS, R does benefit from a community-based ecosystem, and it is correct to think that the answer to many questions during development can be successfully crowd-sourced. This however does not normally include extremely specific situations, where the coder is largely left to her own device.

All in all data ETL can be performed with R, but many times it is awkward at best. The rest of the time it is mostly grueling and slow depending on the exact nature of the task. Many R proponents will readily swear under oath that the same is true of ETL with any alternative to R, but heed our dispassionate advice: if you have the choice between R and Python for ETL, pick Python to walk down the aisle and be forever thankful you did so.

All digital files (including input files, raw and processed data sets, scripts and result files) are made fully available to the reader, in a way which preserves the data structure and the files' hierarchical organization on any computing platform. Paths in adjoined scripts and occasionally in the body of this report are shown using Unix-like formats. However they can be transposed easily to any addressing format of the file system of your choice.

From the top containing folder "*NYC311*", the complete project's file tree is organized as follows:
[…] below means that we omit mention of some intermediate data files, obtained during the preliminary data processing phase. Those files are provided for the record. Their name usually starts with a time-stamp identifying the period to which they refer and ends with __procXX.csv, where XX is a double digit processing sequence identifier.

```
visualCity/
|___ Bibliography/
|___ Data/
|      |___ Geolocation/
|      |      |___ [shape files for NYC ZIP codes and census tract area perimeters]
|      |___ 201x0400_nyc311_raw.csv
|      |___ 201x0400_nyc-crime-map_raw.csv
|      |___ 201x0400_nyc_irs-by-zip.csv
|      |___ [...]
|      |___ nyc_borough-zip.csv
|      |___ nyc311_00083-neighbors-common-border.csv # Ghost zip 00083 processing data
```

```
|      |___ 201x0x00_nyc_whole-data set.csv   # April 201x raw data set
|      |___ 201x0400_nyc_simple-whole-data set.csv # April 201x raw data set
|___ Report/
|___ Scripts/
       |__ 01_nyc311_input-parameters.R # defines basic period parameters and more
|___ Scripts_LDR/
       |__ 02_nyc311_data-prep.R   # clean up of raw data, SRCs' modalities reduction
       |__ 03_nyc311_missing-impute_googlemaps.R
       |      # imputation by direct localization with GoogleMaps' API
       |__ 04_nyc311_calls-by-zip.R # consolidates SRCs' modalities per ZIP
       |__ 05_irs_median-inc-jobless.R # evaluate median income +joblessness per zip
       |__ 06_nypd_data-prep.R # clean up of raw CRC data, reduce to 3 crime modalities,
       |      #+ ZIP imputation by direct localization (GoogleMaps API)
       |__ 07_crimes-by-zip.R # consolidates crime modalities per ZIP
       |__ 08_nyc-data_consolidate-by-zip.R # consolidates SRCs and CRCs per ZIP
       |__ 09a_nyc-zip00083_border-analysis.R # processing of Central Park's ghost zip (1/2)
       |__ 09b_nyc-zip00083_apportionment.R # processing of Central Park's ghost zip (2/2)
       |__ 10a_nyc-zips_find-common-boundaries.R # find ZIP areas included in others
       |__ 10b_nyc-zips_apportion-simplify-data.R # apportion included ZIP areas
       |__ 11_ca-pca-varimax.R # automated analysis using CA, PCA, varimax
       |__ 12a_mca-w-crime-data.R # binify data to conduct MCA on individual data sets
       |__ 12b_mca-time-evolution.R # build consolidated data set including April 2010,
       |      #+ 2014, and 2018, visualize results
       |__ 13a_mca-time-evolution_autonomous-basis.R # visualize independent MCA on 3 data
       |      #+ sets on common plots
       |__ 13b_ mca-time-evolution_common-basis.R # visualize common MCA on 3 data
       |      #+ sets on common plots
       |__ 14_k-means-clustering.R # k-means-clustering with consolidation, visualization
```

2. Data sets

2-1. Terms and conditions of use

All raw data sets used in this project are public and accessible for free under the FOIA. Their use is regulated by the terms and conditions set forth by the governing body responsible for their publication or production. The web pages harboring those terms are:

- http://www1.nyc.gov/home/terms-of-use.page
  for ZIP code centric and time-based NYC311 SRC data
- https://data.cityofnewyork.us/Business/Zip-Code-Boundaries/i8iw-xf4u
  for geometric ZIP code area boundary data
- https://www.irs.gov/statistics
  for ZIP code-centric income tax declaration data made available by the IRS
- https://www.census.gov/topics/income-poverty/income/data/tables/acs.html
  for ZIP code-centric unemployment benefit declared to the IRS

The corresponding data dictionaries are generally made available in Appendix A and Appendix B of this report.

2-2. Data scope and preparation – ETL

Data was generally available from various location on the web, from 2010 onward. We specialized our study to the months of April 2010, 2014 and 2018 in order to be able to handle the corresponding volume of raw data, currently at levels upwards of 250,000 observations of 80 variables per month. Raw files are fully available in cvs formats at `visualCity/Data/`. Census data on population densities per ZIP code area was only available to us for the year 2016 and only for a limited number of ZIP code areas. We therefore could not include that data in any of our data sets.

2-2-1. Duplicates, missings, and imputations

Every downloaded data set was already fully labeled. A rapid inspection of raw data shows that "NA" (non-assigned / not-available) or erroneous values, referred to as "missings", exist, but in such proportion that dealing with them was tractable. As described below, we either imputed, re-imputed, suppressed or researched missings by cross-referencing them between DBs, with the goal of avoiding issues of data bias.

– Service request calls (SRCs) to NYC 311 for NYC's 5 boroughs

The data sets `yyyy0400_nyc311_raw.csv` contain the raw data of NYC SRCs for yyyy={2010,2014,2018} as downloaded from *NYC Open Data*. That includes the call's object (description), date, time, ZIP codes and/or location (in several forms) of the reported matter and other less relevant information. We checked that data sets contains SRCs (heretofore referred to as "dupes") from different callers with the exact same object, i.e. calls are logged from different callers for the same matter . Tracking down dupes is inherently complex and we did not attempt it. More importantly, our study is concerned with people's spontaneous and independent tendency to call NYC 311, about aspects of their urban environment, which are important to them. In that sense dupes need not be eliminated; they are significant and represent a natural weighting for the data set's observations. This will naturally influence observations' weights as represented later by marginals (row sums) in frequency tables.

Raw (unfiltered) data characteristics are shown in Table 1 for the April 2010, 2014 and 2018 data sets.

Period	Raw data's observ. number	Observ. with missing ZIP	Observ. missing all location info	SRCs' raw modalities #	Unique ZIP codes
April 2010	158,398	12,068	2,976	175	312
April 2014	81,645	3,206	2,740	170	278
April 2015	101,890	4,231	3,069	178	260
April 2018	199,840	7,581	2,485	197	293

***Table 1**: Summary table of salient missings and other characteristics for raw NYC 311 SRCs data sets before data cleaning). SRCs' modalities are available in the 2 files: `Report/yyyy0400_nyc311_proc01_modalities.csv` where yyyy is the 4-digit year.*

Figures 2 and 3 below represent missings for the period April 2014.

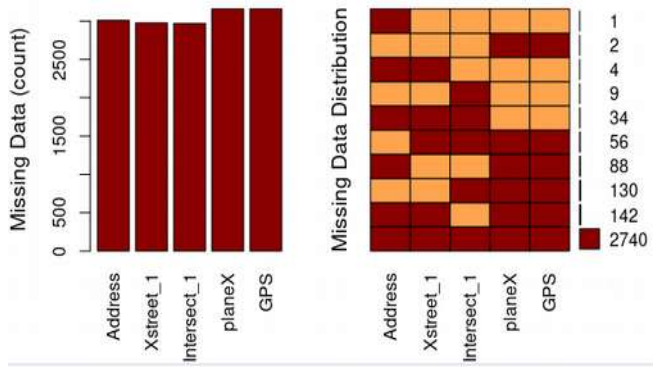


***Fig. 2**: Analysis of missings for April 2014 on the 3206 raw observations missing a ZIP code.*

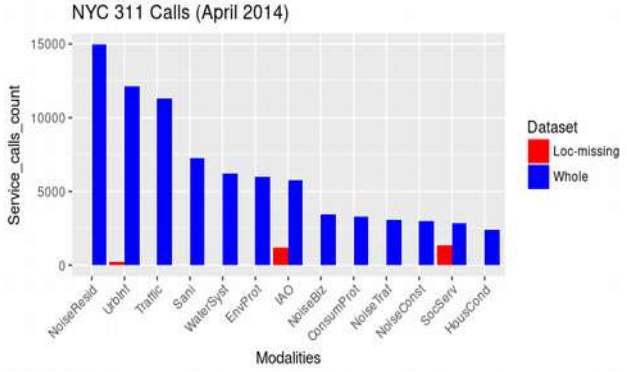


***Fig. 3**: NYC 311 SRCs' modality distribution for the whole data set (blue) and for observations missing all location information.*

As observed from Figure 2 above, during the April 2014 period, 2740 observations or 3.4% of all observations, and 85.5% of the 3206 observations missing a ZIP code have no other geographic locator. Those observations cannot be attributed to any ZIP code and are therefore useless. Figure 3 compares the service request calls' modality distributions for observations missing all location information (including a ZIP code and denoted "*loc-missing*") and the whole data set. It is readily apparent that simply eliminating "*Loc-missing*" observations would disrupt our analysis in terms of the *SocServ* modality, while for other modalities the effect would be negligible.

For that reason, we proceeded to impute a ZIP code to the 466 SRCs missing it in 2014, but not included in the *Loc-missing* subset of missings. In practice those observations miss a ZIP code but are nevertheless endowed with some other geolocation information:

- an address, and/or
- cross-streets in the form of (Xstreet_1,Xstreet_2), and/or
- a cross-road in the form of (Intersect_1,Intersect_2), and/or
- partial information pertaining to at least two of the above, and/or
- planar (Euclidian) coordinates (planeX, planeY), and/or
- GPS coordinates (latitude and longitude)

Imputation was done by fully implementing automated requests to GoogleMaps, through its API, in R, for each one of the aforementioned cases. As a result more than 97% of all 466 observations missing a ZIP code could be imputed for the April 2014 data set. The rest including the *Loc-missing* subset of observations were given the bogus ZIP code "99999" to be uses later as a supplementary observation.

As there is no structural difference between the April 2014 and data sets covering other periods (2010, 2015, 2018), graphical analysis results for missings arere only shown for April 2014. From Table 1, in April 2015, 3069 observations or 3.0% of all observations, and 72.5% of all observations missing a ZIP code have no other geographic locator. Here again we can treat missings following the same pattern and with a similar success rate as before. Hereafter April 2015 is only mentioned as a model example for expedient comparisons.

– Crime report calls to 911 (NYPD logged CRCs) for NYC's 5 boroughs

Crimes are reported according to 3 general categories, which coincide with the crime modalities used in our analysis. In decreasing order of severity, they are: **felonies**, **misdemeanors**, and **violations**. . They are described and instances listed in Appendix B per the NYPD's DB.

Data made publicly available by NYDP is completely devoid of ZIP information. However it does include planar localization and regular GPS coordinates. Because of the large amount of data involved in this study (close to 80,000 criminal observations) and of Google's imposed limitation on the number of queries (2500/day/account, as of 2018.04.30) , relying on our Google Maps API's implementation to impute a ZIP code to each crime was not deemed practical. We therefore developed two original algorithms to determine the ZIP code of each NYPD crime observation based on its planar (Cartesian) coordinates. Table 2 exhibits some of the salient counts in this area.

April 2014	**Felony**	**Misdemeanor**	**Violation**	Total
non-missings	11,327	22,094	4,784	38,205
missings	481	985	64	1,530
Total	11,808	23,079	4,848	39,735

April 2015	**Felony**	**Misdemeanor**	**Violation**	Total
non-missings	11,669	22,080	5,010	38,759
missings	193	473	11	677
Total	11,862	22,553	5,021	39,436

***Table 2**: Summary of misssings after imputation for the NYPD's crime data sets in NYC*

The first algorithm to be developed was based on nearest neighbor topological distance. It uses previously compiled ZIP code areas with planar and/or GPS coordinates for SRCs to NYC 311. The ZIP code of the 311 SRC closest in space to a crime's GPS or planar coordinates is imputed to the crime. This method is approximate and yield mixed results.

The second algorithm is exact (and somewhat complex) and yields excellent results. It determines the ZIP code of every crime observation based on its planar coordinates and shape-formatted ZIP boundaries mapping data, downloaded from the *NYC Open Data* repository and made available to the reader under `Data/Geolocation/`.

The latter algorithm is general and is implemented in the form of a function, `whichBoxF()`, available at `Scripts_LDR/06_nypd_data-prep.R`. Its reaches its imputation target in more than 96% of all recorded observations. The rest, i.e. less than 4%, falls in the *missings* category and kept in supplementary observation with imputed bogus ZIP code "99999". Tables 2 below summarizes missing ZIP code "99999" imputation for crime data collected by NYPD in April 2014 and April 2015. A Chi square test of the NYPD crime data sets' missings show that there is a significant association between missings and crime modalities. Simply suppressing missings would introduce a bias in the distribution.

2-2-2. SRCs' modality dimensional reduction

Service Request Calls' modality dimensional reduction was conducted by applying filters tailored to the semantics of the raw data's two columns: "Complaint", and "Descriptor".

The reduced modalities data sets exhibit 13 modalities down from 170 and 178 (in Table 1, for April 2014 and April 2015 respectively) according to the description and distribution of Table 3. Noise related complaints remain the first reason for SRCs to 311 in NYC, with overall frequencies in noise related calls of 31.1% and 31.5% in 2014.and 2015 respectively.

Service request calls' modalities	Modality description	Service request call frequencies		Change in rank from 2014 to 2015
		April 2014	April 2015	
NoiseResid	Residential Noise	19.00%	17.50%	–
UrbInf	Urban Infrastructure	15.00%	13.40%	↘
Traffic	Traffic related Issues	14.30%	17.20%	↗
Sani	Unsanitary Conditions	9.20%	10.50%	–
WaterSyst	Water Systems	7.80%	7.60%	–
EnvProt	Environmental Protection	7.60%	5.90%	–
IAO	Inspect, Audit, Order	5.80%	5.20%	↘
NoiseBiz	Commercial Noise	4.40%	4.90%	↘
ConsumProt	Comsumer Protection	4.20%	3.40%	↘
NoiseTraf	Traffic Noise	3.90%	5.40%	↗↗
NoiseConst	Construction Noise	3.80%	3.70%	↗
HousCond	Housing Conditions	3.10%	3.40%	–
SocServ	Social Services	1.90%	1.90%	–
Total number of SRCs		78825	98649	↗↗

***Table 2**: SRCs' consolidated modalities after dimensional reduction. The right most column indicates changes in modality ranking from 2014 to 2015.*

Table 2 is based on data after ZIP cleaning and missings imputation. SRC modality ranking change show that the perceived (and perhaps also real) traffic noise related SRCs increased markedly between April 2014 and April 2015.

2-2-3. ZIP code cleaning

At this data preparation stage, the data consists of a mixture of correctly formed and ill-formed ZIP code fields for each observation. An ill-formed ZIP code may be a code, which either does not have exactly 5 digits, or does not exists officially, or is otherwise not consistently found in US federal or municipal DBs.

To easily associate ZIP codes and borough, we include a list of 200 ZIP codes and corresponding boroughs in Appendix C.

For our purposes, ill-formed ZIPs include ZIP+4 codes of the form 11355-1024, where the last four digits identify a geographic segment or a PO box within the five-digit ZIP delivery area. In those cases we simply suppress string characters from position 6 to the end.

Inadmissible ZIP codes also include ghost ZIP codes. One of them appears in our DBs as "00083". The NYC 311 service request call data set includes it along with surrounding and overlapping ZIP codes. So do the NYPD's crime DB, and the topological ZIP code area boundary DB also found in the NYC Open Data repository. Within the NYC area it designates very precisely the Central Park area in Manhattan. But because it overlaps with other official ZIP code areas surrounding it, observations identified by that ZIP code should be instead apportioned to neighboring ZIP code areas. Figure 4a reveals the Zip mapping in that area, showing official ZIP code areas boundaries mapping Central Park in Manhattan. Surrounding ZIP codes are 10019, 10022, 10065, 10023, 10021, 10075, 10028, 10024, 10128, 10025, 10029, and 10026.

The use of ZIP code 00083 is incompatible with IRS and other federal agencies' DBs (such as that in charge of census). To overcome that difficulty, we calculated the common boundaries between the 00083 ZIP code area boundary and surrounding ZIP areas boundaries based on the Universal Transverse Mercator (UTM) coordinates system.

Our goal is to apportion observations attributed to ZIP code 00083 to surrounding ZIP codes areas proportionally to the lengths of the boundaries they share, in a way which should be modality-neutral.

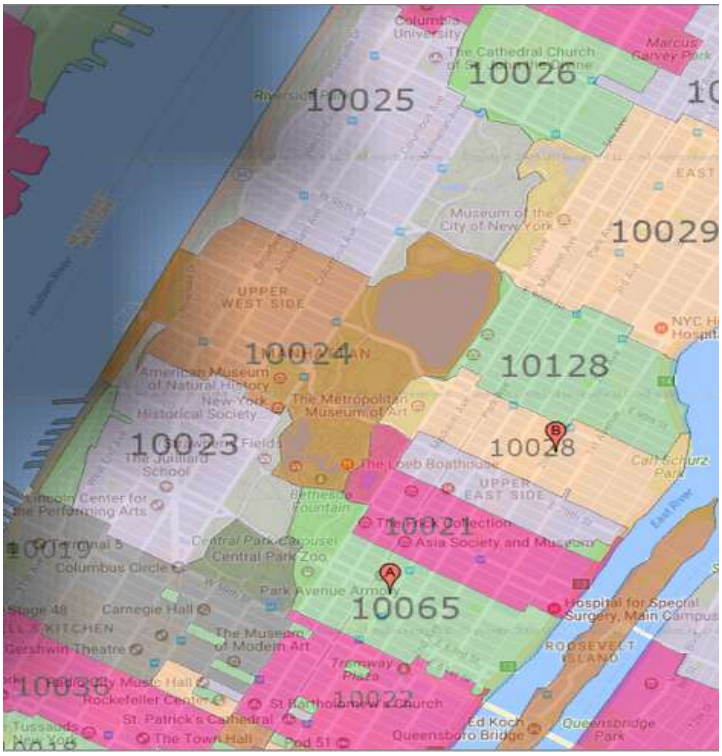


***Fig. 4a**: Detail of the ZIP code area map of Manhattan, showing how neighboring ZIP code areas pave Central Park piece-wise.*

ZIP code	Common boundary length (ft)	Common boundary length proportion (%)
00083	32,710.8	100.0
10019	2,651.4	8.1
10022	259.2	0.8
10065	2,341.4	7.2
10023	4,864.4	14.9
10021	2,150.5	6.6
10075	833.0	2.5
10028	1,889.7	5.8
10024	3,748.7	11.5
10128	2,371.1	7.2
10025	5,031.3	15.4
10029	3,749.0	11.5
10026	2,821.2	8.6

***Table 4**: 00083 ghost ZIP code area shared boundary analysis.*

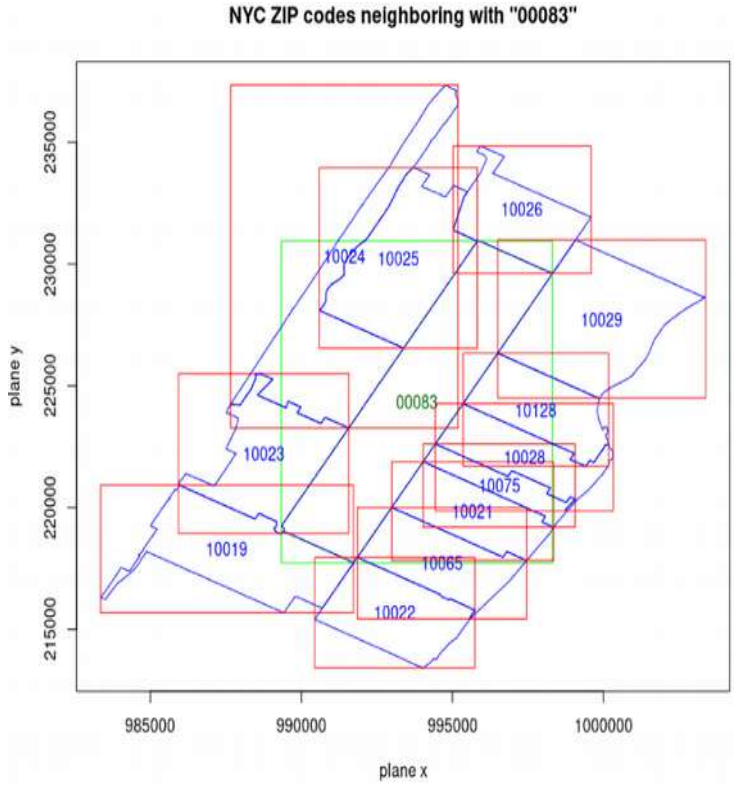


***Fig.4b**: New York City government ZIP code limits mapping Central Park in Manhattan and resorting to ghost ZIP code 00083 (at center in green).*

Figure 4b represents the UTM topological mapping, and Table 4 shows the computed proportion of common boundary lengths between Central Park's 00083 ghost ZIP and surrounding ZIP code areas. The algorithm developed can operate on arbitrary sets of ZIP codes.

After correcting for ill-formed ZIP codes, ghost ZIP codes, ZIP codes with zero surface area (i.e. corresponding to PO boxes), ZIP codes situated outside NYC's 5-borough area, we observed a little over 200 unique ZIP codes in our data sets (year in year out).

Finally we let a limited number of ZIP codes areas be absorbed by their "main neighbor", according to the following rationale: whenever at least 75% of any two given ZIP codes' boundaries coincide, we apportioned the observations counts of the ZIP code whose area had the shorter overall boundary length to the neighboring ZIP code area with the longer boundary length. We are aware that in a limited number of very specific cases apportionment based solely on that criterion is not optimal.

The starting point for our statistical analysis of section 3 are the three data sets located at:

```
visualCity/
      |___ Data/
             |___ 20100400_nyc_simple-whole-data-set.csv
             |___ 20140400_nyc_simple-whole-data-set.csv
             |___ 20180400_nyc_simple-whole-data-set.csv
```

We noted a few cases (fewer than 30 observations per data set) of missing response variables ("`medianInc`" and "`jlBenefit`") in each data set. This means that the Internal Revenue Service (IRS) chose not to make the corresponding ZIP code area's tax return statistical data public. This should not pose a problem for the coming clustering analysis (classification) , but obviously does so in any regression or decision-tree-like approach requiring any of those two response variables as control.

So the reader may comfortably associate ZIP codes with NYC boroughs, we provide a list of more than 200 ZIP codes and their corresponding boroughs in Appendix C.

3. Multi-Variate Analysis

3-1. Scope of observations

3-1-1. Aggregation scale of observations

We choose to conduct the analysis presented hereafter at ZIP code area scale. Results are commented at that scale as well as at borough area scale. This is mainly for the convenience of city management and urban administration are ZIP core area boundaries almost always follow natural urban landscape boundaries such as streets, avenues, water ways or park limits. ZIP code areas however may well represent heterogeneous populations and/or urban geography. They are also subject to change, for no other purpose than to be administratively convenient for the US postal service. We could as or more easily conduct the same analyses not at ZIP code, but at census tract level, at building block levels or even (on the basis of isolated events) at GPS coordinates level.

At GPS coordinate ("event") scale for instance, every logged call, every reported crime, in general every included event would constitute an observational data point. This in turn would force us to debate another important aspect of our analysis. In the context of clustering analysis, and as event would be the new individual data points, how appropriate would weighted or unweighted event approaches be in order to reveal the semantics of urban dynamics ? In short, a weighted event approach treats all clusters equally, while an unweighted one considers that cluster classes are more or less important as a function of their memberships (i.e. their cardinality). In general, unweighted approaches are preferred unless there is reason to believe that observations should have different weights; e.g., perhaps because classes of objects were unevenly sampled, etc ...

In this exploratory work we chose to consider our observation data as ZIP code scale aggregates, in order to avoid the above discussion, due to the fact that our categorical variables and their modalities are largely dissimilar in nature. A crime or an

offense are not generally reported arbitrarily, i.e. following the whim or current state of mind of the persons reporting them, but rather because, being a breach of social contract, failure to report may in itself be a punishable offense. By contrast a service request call to 311 by a NYC resident or visitor may go unreported without consequence for the witnesses. In all likelihood, certain city dwellers are less prone than others to report urban issues as they perceive them; visitors arguably much less still. This is turn may not be ascribed exclusively to the psychological make-up of any individual, but probably also to a large number of external factors which influence and determine at least in part any potential caller's decision to call. SRCs are in this sense arbitrary and are (at least from the perspective of an uninformed analyst) placed at will. That however constitutes an event sampling mechanism completely different from that of CRCs. It would likely introduce a bias when performing MVA at event scale. In our opinion and for the sake of simplicity, this was ample justification for considering our analysis at ZIP code area scale.

3-1-2. Low observation counts

Our first approach was to consider the contingency table made of the *NYC 311 SRCs* categorical variable's 13 modalities and 200+ zip codes seen as the modalities of a second categorical variable we name *Location*.

Among the sorted zip codes, the last one, "99999", will either be overlooked or be treated as a supplementary observation.

We identify between 20 and 30 zip codes with row marginals smaller than 5/(sum of calls), where, e.g. for April 2014, the total number of calls so far retained in our analysis was about 78,700. We suppress those ZIP codes from our contingency table, on the grounds of they representing less than 0.2% of monthly SRCs (see footnote[viii]). The resulting table for April 2014 is made of 181 zip codes (row labels, row index *i*) and 13 SRC modalities (column labels, column index *j*).

Next we identify table cells where low frequency and (simultaneously) high contributions to the χ^2-statistic value for the test of association of the two categorical variables may perturb the subsequent analysis. We define as low cell count or low frequency any contingency table cell count smaller than 5. For every data set there are between 300 and 500 such cells. Based on the chi-square-test statistic:

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(Count_{obs} - Count_{exp}\right)^2}{Count_{exp}}$$

we calculated the contribution of every low frequency cell to the overall χ^2 statistic value and found that for low frequency cells: (i) no contribution exceeds 1%, and (ii) only 1 contributions exceed 0.1%, for a 2-sided χ^2 test statistics of 43,338. As a result the Pearson chi-square test for significant association (dependence) between row & column categories is deemed appropriate. It leads to the clear rejection of the null hypothesis, with a p-value of the order of 10^{-4}:

H_0: "In the population, the two categorical variables are independent."

The above p-value was computed from Monte-Carlo simulations with 10,000 replicates.

Inspecting marginals, we see that SRCs' modalities with lowest weight across zip codes are:
"SocServ" ($f_j \approx 0.019$ for j=11), followed by HousCond ($f_j \approx 0.030$ for j=1), and "NoiseConst" ($f_j \approx 0.038$ for j=4).

3-2. Principal Components Analysis (PCA) and Correspondence Analysis (CA)

3-2-1. PCA

From the contingency table made of the April 2014 SRC categorical variable's 13 modalities and 202 ZIP codes, we build a conditional frequency matrix, which we appropriately center based on a cloud centroid (of column marginals) with embedded χ^2 metric. We perform a PCA [1] on that matrix excluding ZIP code "99999" as well as 21 other individual ZIP codes whose marginal row counts are smaller than or equal to 5. 180 individuals are left. We first include "10463" and then repeat the analysis considering it as a supplementary observation. The number of significant dimensions is 3, based on the criterion that the total explained inertia be at least 70%. Results are graphically summarized in Figures 5a, 5b and 5c.

viii *A χ^2-test of independence on the small contingency table made of ZIP codes to be suppressed and their RFCs' modalities led us to reject the null hypothesis of independence. To that end, data was reduced so no zero valued marginals could perturb the test.*

When included in the analysis (as in Fig. 5a), ZIP code "10463" stands out as the biggest individual contributor to the construction of the 3 first dimensions with 17%, 22%, and a whopping 52% for PC1, PC2 and PC3 respectively. The ZIP code area roughly represents a one kilometer radius in the Bronx, known as Riverdale. Topologically neighboring ZIP areas are: 10467, 10468, 10471.

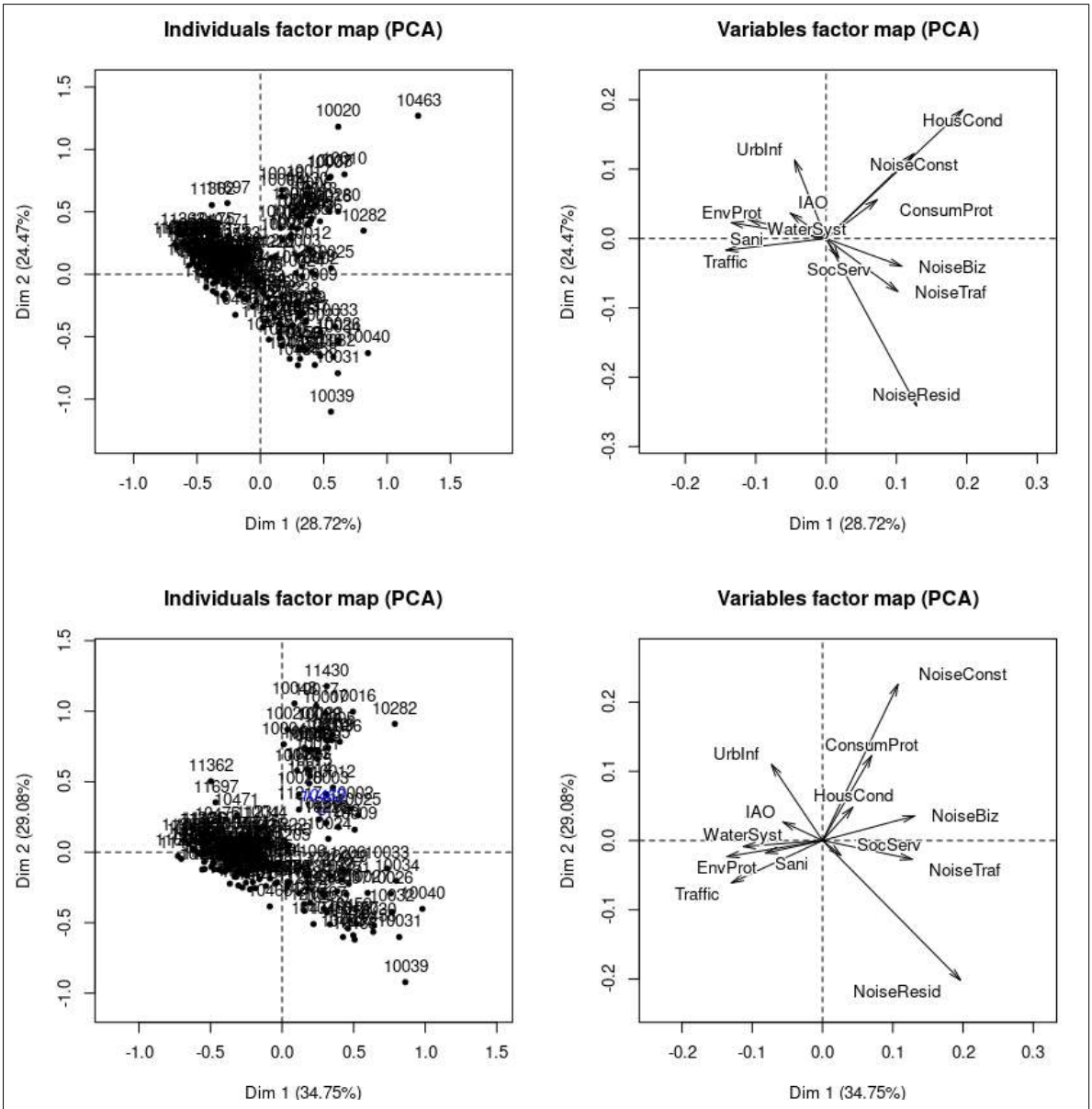


***Fig. 5a**: Weighted observations' and variables' first factorial plane (PC1-2) scatter plots obtained by PCA: (top) including and (bottom) excluding ZIP code "10463". The effect of the χ^2 metric is incorporated in the projections. Bogus ZIP code "99999" is not included and outlier "10463" when not included is a supplementary individual represented in blue on the left plot.*

Riverdale has one of the highest population density in NYC with more than 30,000 housing units and more than 18,000 registered inhabitants per square kilometer. Understandably `HousCond` and other SRCs to NYC 311 are disproportionately large in Riverdale, when compared to other NYC areas.

Besides a noticeable change in cloud shape, a pronounced change takes place when we consider "10463" as a supplementary individual. It concerns principally the variable `HousCond`, whose:

- quality of representation in the first 3 dimensions (PC1, PC2, PC3), and

- contributions to the construction of dimensions

both plummet. Meanwhile the contributions and quality of representation of the other two main variables `NoiseResid` and `NoiseConst` are somewhat redistributed among dimensions or in some cases increased: e.g. for `NoiseResid` $\sum_{\alpha=1,2,3} \cos^2_\alpha$ goes from 0.94 to 0.98.

The first factorial plane (PC1-2) registers an increase in inertia explanatory power (from 54% to 64%) – Fig. 5a. Meanwhile PC2-3 and PC1-3 register a decrease from 48% to 43% and from 44 % to 37% respectively (see Fig. 5b and Fig. 5c).

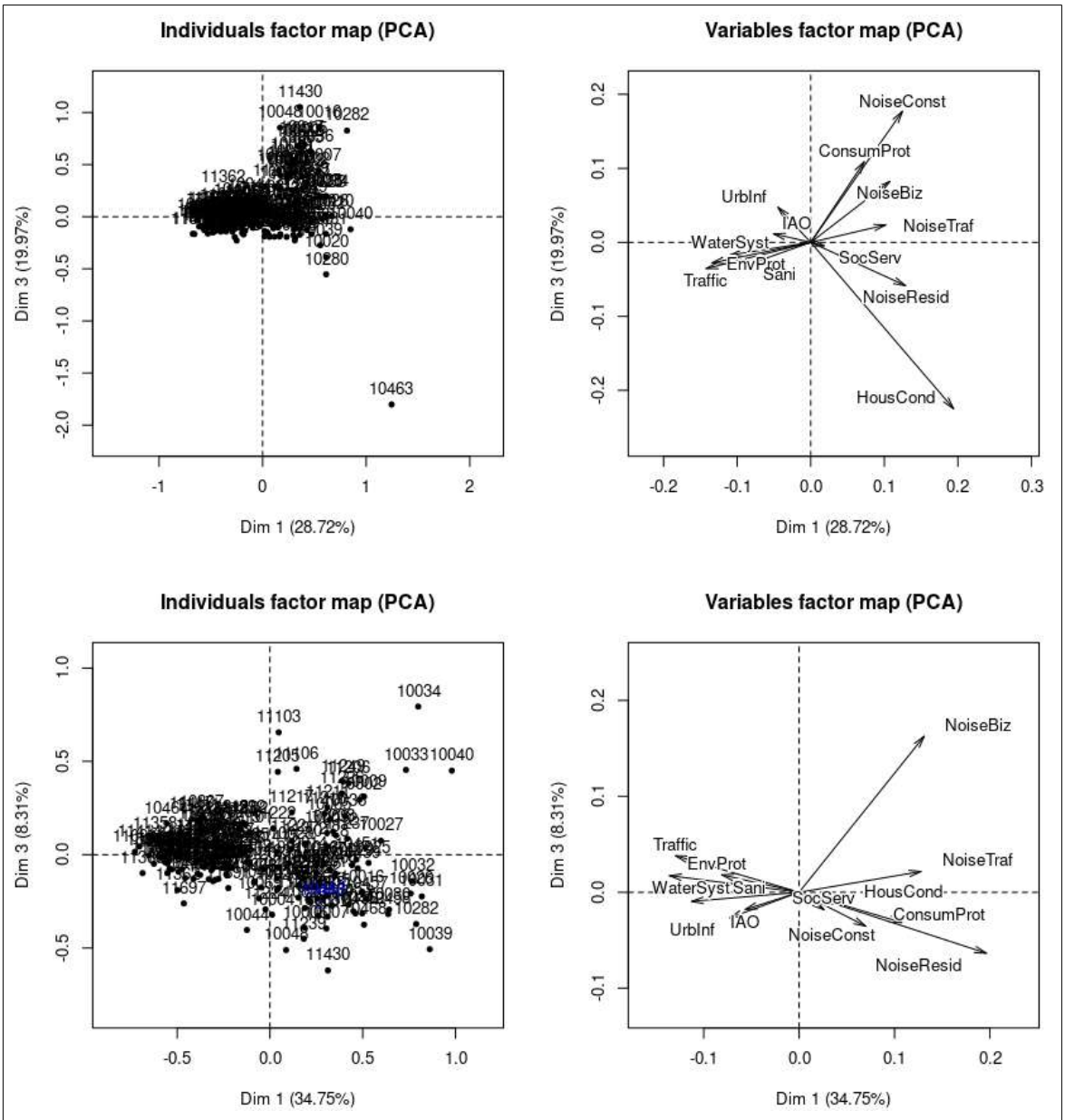


***Fig. 5b**: Weighted observations' and variables' second factorial plane (PC1-3) scatter plots obtained by PCA: (top) including and (bottom) excluding ZIP code "10463". The effect of the χ^2 metric is incorporated in the projections.*

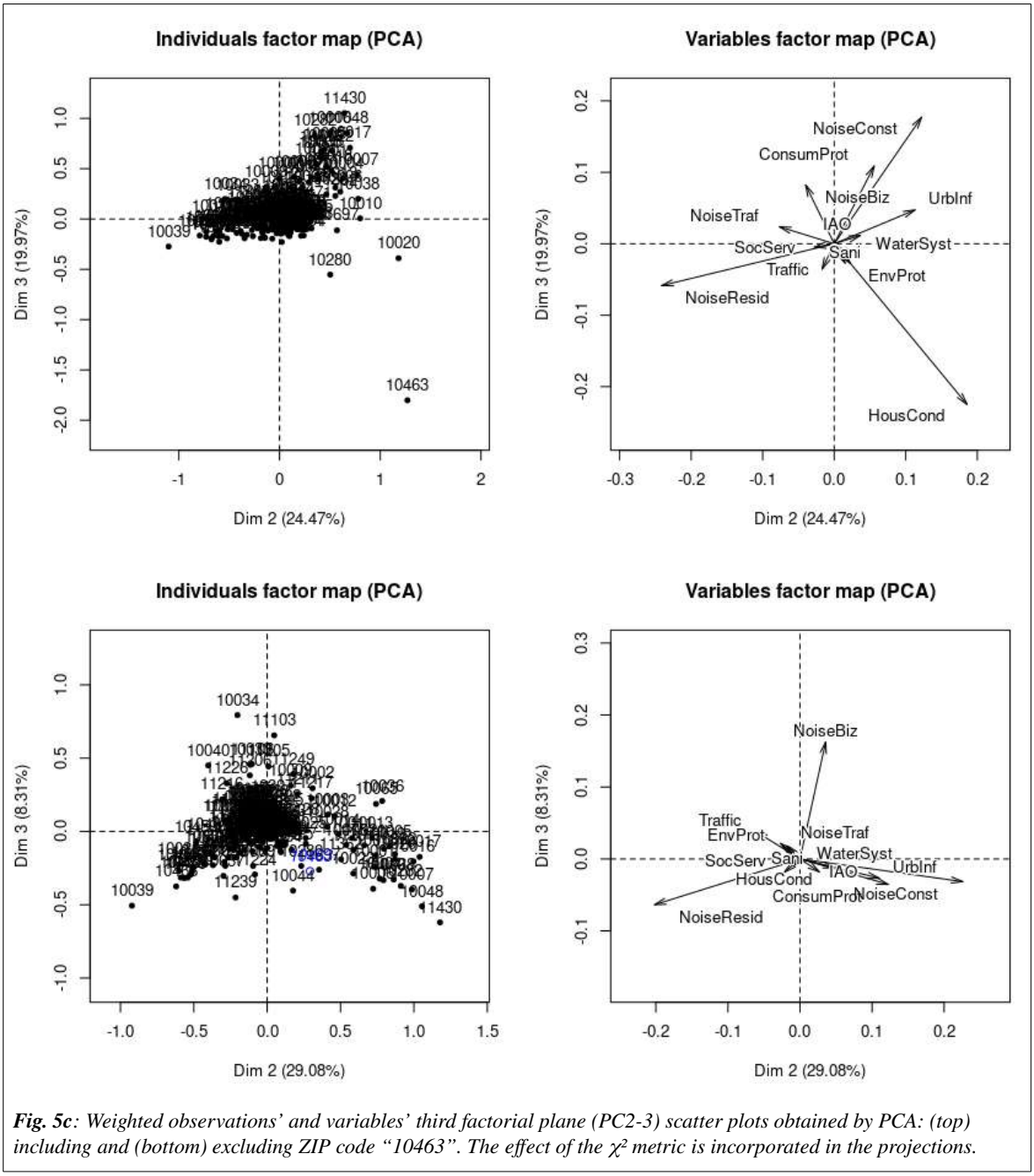


***Fig. 5c**: Weighted observations' and variables' third factorial plane (PC2-3) scatter plots obtained by PCA: (top) including and (bottom) excluding ZIP code "10463". The effect of the χ^2 metric is incorporated in the projections.*

We observe on the variable factor maps, considering "10463" as supplementary individual, that:

- `NoiseConst` and `ConsumProt` appear to be largely collinear, capturing together more than 32% of the data's dynamics. We decide not to join them however as no satisfactory justification was found to explain the apparent correlation.
- `EnvProt` and `Sani` are consistently collinear and may be merged into a new feature called `EPsani`, capturing more than 8% of the data dynamics.
- `EnvProt` and `WaterSyst` appear strongly correlated in planes PC1-2 and PC1-3, but anti-correlated in PC2-3. `WaterSyst` however exhibits a relatively poor Inertia Explanatory Power (see Table 5 below) with IEP < 5% for the retained significant dimensions. This justifies eliminating `WaterSyst` in an effort to decrease dimensionality.
- `NoiseConst` and `HousCond` are largely collinear in planes PC1-2 and PC1-3, but anti-correlated in PC2-3. As previously for `WaterSyst`, `HousCond` being a weak variable with IEP < 5% for the retained significant dimensions, justifies doing away with `HousCond`.

- *IAO* and *SocServ* seem to play a negligible role in explaining variance and may be altogether dispensed with.

Among the largest contributors to the construction of the 3 first principal directions, we highlight the fact that *NoiseResid*, *NoiseConst*, *Traffic*, *NoiseTraf*, *ConsumProt* and *EnvProt* are all best represented in the first factorial plane (PC1-2).

By contrast *NoiseBiz* is best represented by the second factorial plane (PC1-3, see figure 5b above), where it plays a dominant role in the construction of the 3rd dimension, PC3.

3-2-2. CA

We conducted a Correspondence Analysis (CA) with row marginals as row profile's weights, thereby incorporating the χ^2 metric effect into the row-profile cloud projection.

Distances *between identically colored points* are distances in the χ^2 sense to correct for the relative scarcity of factors. A red point (column profile) is a barycenter for the blue points (row profiles) expressing that column modality, weighted by said column, and vice versa.

Differently colored points may appear close, but no conclusion can be drawn from that apparent proximity on the graph. On the other hand, identically colored points, which are close together, do have similar profiles.

Next Tables 5a and 5b below exhibit the inertia explanatory power (IEP) for each SRC's modality, alternately considering all dimensions and only significant dimensions before and after feature selection and dimensionality reduction.

SRCs' modalities	13D-IEP(%)	3D-IEP (%)
HousCond	4.0	1.9
Sani	4.2	2.3
NoiseResid	19.0	26.0
NoiseConst	15.4	19.6
NoiseBiz	10.4	13.9
UrbInf	6.0	5.7
Traffic	8.9	6.9
NoiseTraf	6.3	5.5
WaterSyst	6.2	4.0
ConsumProt	7.1	6.5
SocServ	1.6	0.5
IAO	3.0	1.3
EnvProt	7.9	6.0

SRCs' modalities	8D-IEP(%)	2D-IEP (%)
NoiseResid	21.2	27.0
NoiseConst	19.3	23.9
NoiseBiz	12.1	6.3
UrbInf	8.6	8.1
Traffic	11.7	10.1
NoiseTraf	7.2	5.8
ConsumProt	8.8	7.5
Epsani	11.2	11.3

***Table 5a (left)** features factors' inertia explanatory power, **before** feature selection, over all 13 dimensions and for 3 significant dimensions (shaded cells have IEP>5%).*
***Table 5b (right)** shows the same after feature selection and dimensionality reduction.*

In Fig. 6, **row (blue)** and **column (red)** profiles are projected together as biplots, after feature selection and extraction, considering ZIP code "10463" as a supplementary observation.

The scree plot above reveals 2 significant dimensions with eigenvalues (in decreasing order of inertia representation): 0.16, 0.15 for a total explained variance of almost 73%.

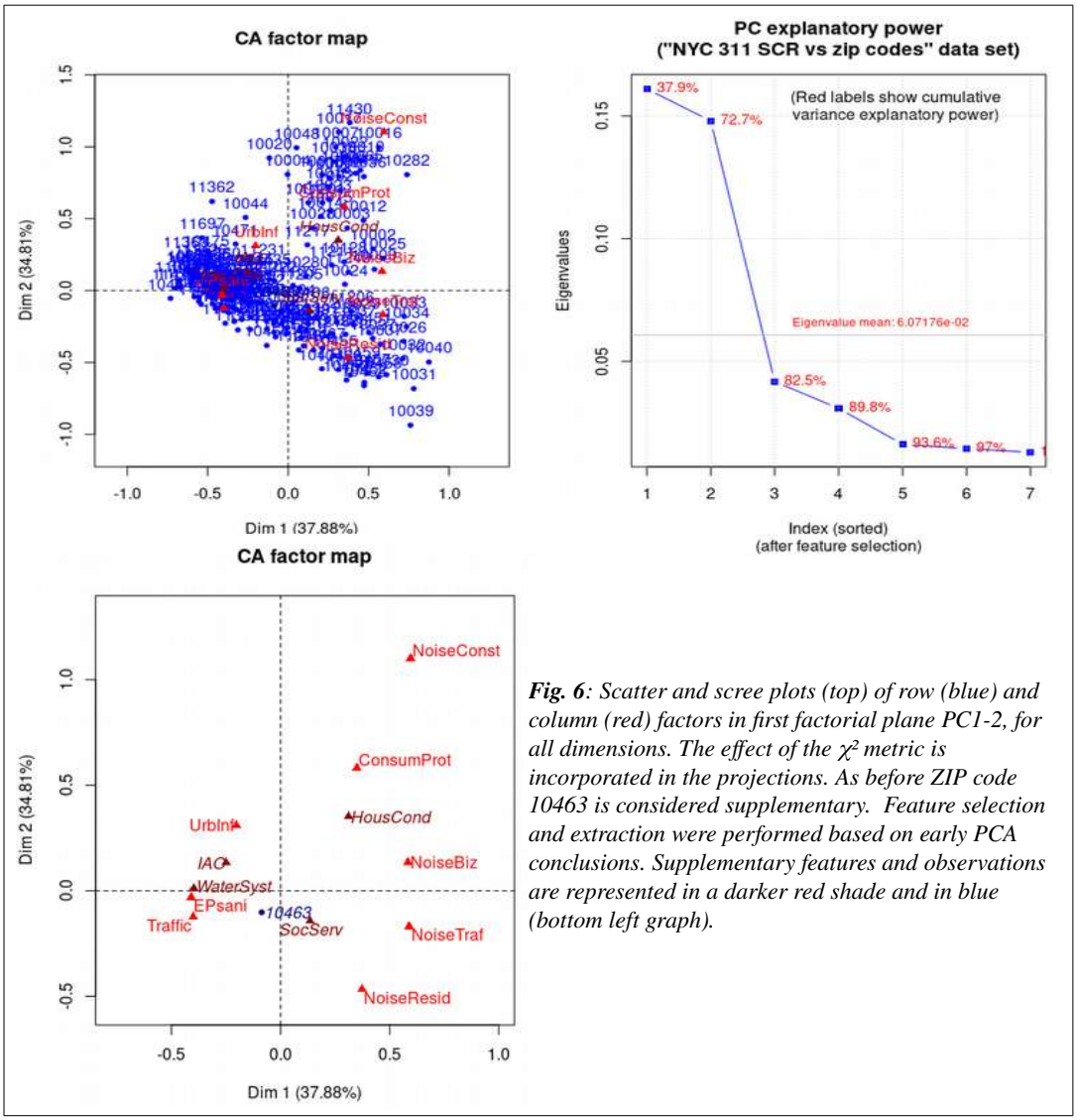


***Fig. 6**: Scatter and scree plots (top) of row (blue) and column (red) factors in first factorial plane PC1-2, for all dimensions. The effect of the χ^2 metric is incorporated in the projections. As before ZIP code 10463 is considered supplementary. Feature selection and extraction were performed based on early PCA conclusions. Supplementary features and observations are represented in a darker red shade and in blue (bottom left graph).*

Figures 7 and 8 exhibit variable and individual projections in PC1-2 after feature selection and extraction.

Figure 7 shows that `NoiseResid` and `UrbInf` are anticorrelated. Areas of high incidence for `NoiseResid` SRCs exhibit low incidence of `UrbInf` related calls, as if populations beset by residential noise from neighbors were less prone than others to complain about surrounding urban infrastructure in their areas. The inverse may also hold as we make no hypotheses about a tie of causality between the two SRC's modalities. Also worthy of note is the fact that `NoiseResid` and `NoiseConst` are very weakly correlated.

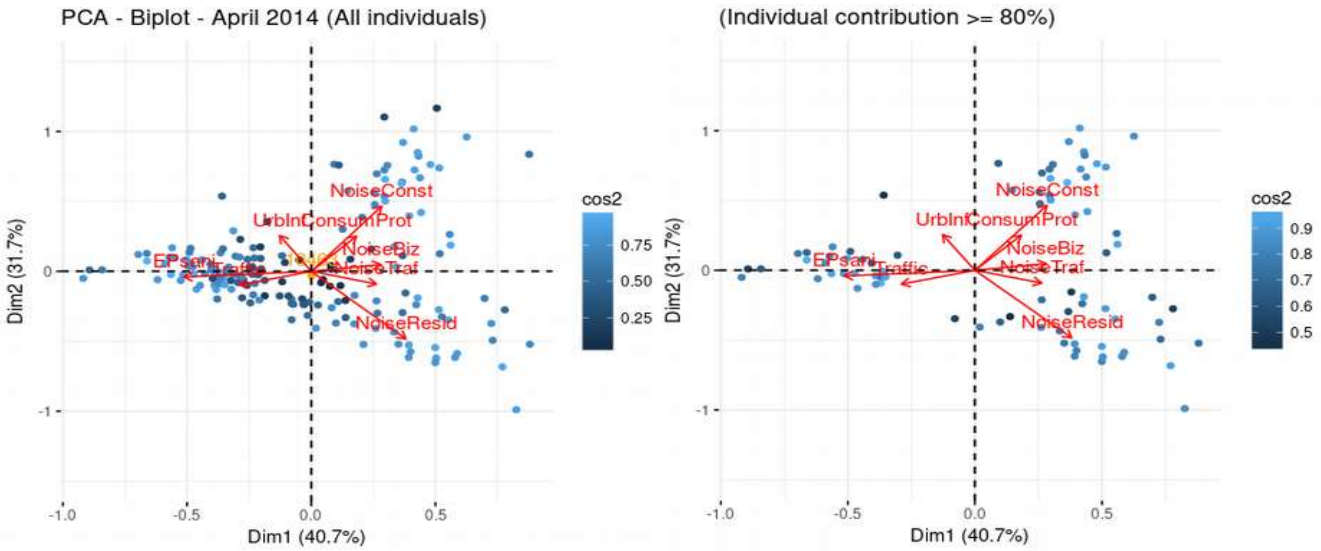


***Fig. 7**: Biplots in the 1st factorial plane PC1-2 for April 2014. The effect of the χ^2 metric is incorporated in the projections. ZIPs codes 99999 and 10463 are included (in gold) and both feature selection and extraction were performed based on previous PCA observations. Both figures show color-coded individuals, according to their quality of representation (cos2). The right plot is restricted to individuals whose contributions to the construction of the first 2 dimension is greater than 80%. Arrows represent direction of growth for features.*

From the borough-based color-coded visualization of scores in Figure 8, one further notes that:

- Manhattan's make-up (dark blue dots) is heterogeneous appears characterized by `NoiseConst`, `NoiseBiz`, `NoiseTraf`, `NoiseResid`, and `ConsumProt`,
- Most of Queens (golden dots), and part of the Bronx (red dots) are consistent with higher incidences of `EPsani`, `Traffic`, and `UrbInf` related complaints,
- Staten Island (cyan dots) appears fully characterized by a majority of complaints under `Epsani`, and `Traffic`,
- In addition to the above, the Bronx (red dots) is also characterized by `NoiseResid` related complaints,
- Brooklyn's ZIP codes projections (green dots) are relatively difficult to interpret as they seems to simultaneously extend in all 4 quadrant, and is therefore representative as a borough of all SRCs' features and type of complaints.

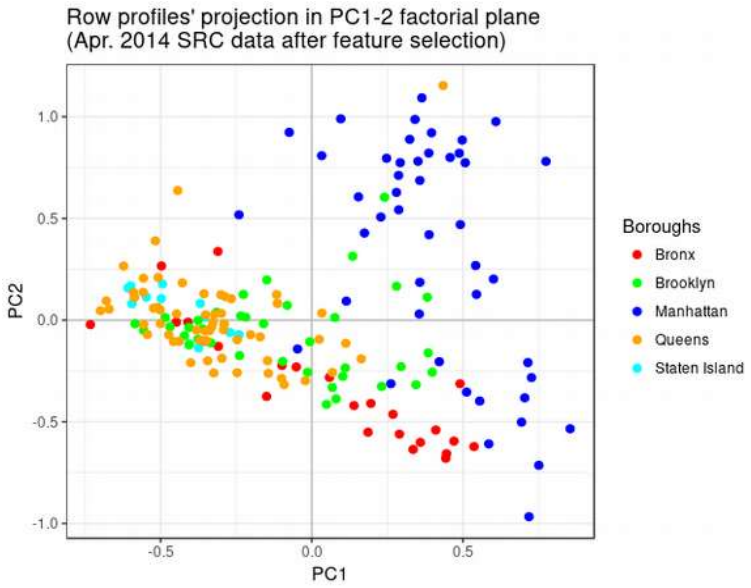


***Fig. 8**: First factorial plane map of individual ZIP codes for the period April 2014, color coded according to the NYC borough to which they belong.*

Table 6 (below) summarizes SRC individuals' explanatory power per borough for all dimensions and for only the first factorial plane (i.e. for the 2 significant dimensions).

Borough	number of ZIP codes	IEP all_dim (%)	IEP 2 signif dim (%)
Bronx	24	13.9	10.8
Brooklyn	38	16.2	9.8
Manhattan	46	46.2	37.1
Queens	59	17.2	9.3
Staten Isl.	12	6.0	4.4

***Table 6**: Inertia explanatory power by individual ZIP codes grouped by borough, computed over all dimensions (3rd column) and over the significant dimensions (4th column).*

The PCA based initial exploration of NYC's SRCs is almost concluded with a topographical map (Figure 9, right) of row individuals (i.e. ZIP codes), color-coded according to the position of their projection in the PC1-2 factorial plane, following Figure 7. Dot colors represent row profiles' (i.e. individual ZIPs') projections in the four PC1-2 quadrants:

1st quadrant (**orchid**),
2nd quadrant (**green**),
3rd quadrant (**tan**),
4th quadrant (**red**).

As previously noted Brooklyn covers the complete range of SCRs modalities, as shown by the fact that the borough contains dots of all four colors.

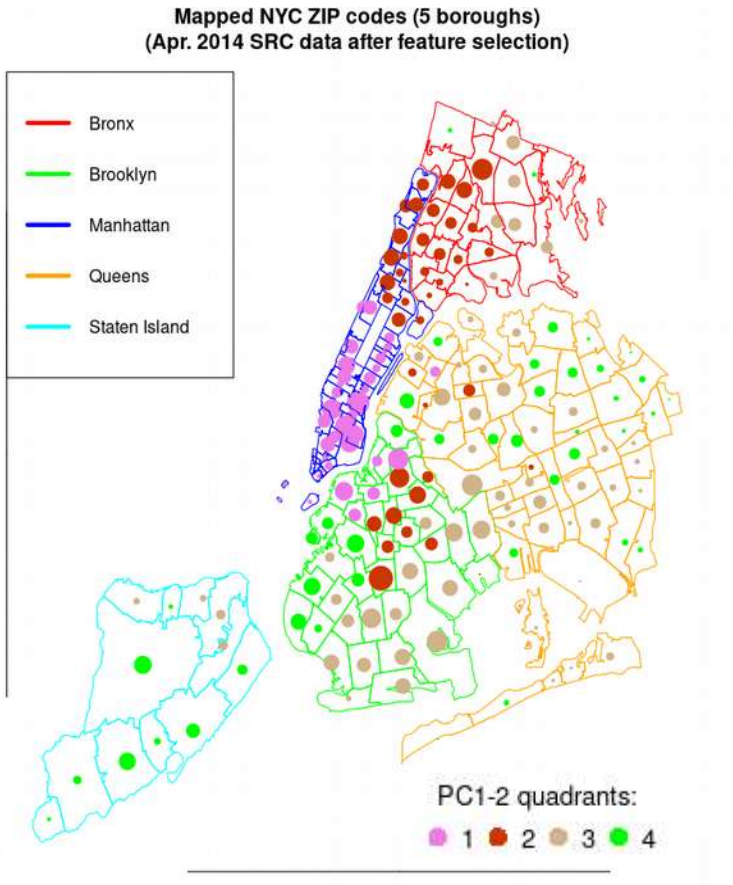


***Figure 9**: Topographical representation of ZIP codes' projection quadrant in the first factorial plane (per Fig. 8) for the period April 2014. Dot sizes are proportional to the number of SRCs in a given ZIP code area*

3-2-3. Varimax applied to PCA / CA results – Latent factor analysis

R's *varimax()* method [2] to find latent concepts, is a simple factor-oriented (i.e. *column-oriented*) structure rotation designed to maximize the sum of column-wise variances of the squared loadings, that is, the squared correlations between variables and factors. The approach aims at interpreting results in the reduced sup-space of the (in our case) two rotated significant directions. It does not *generally* preserve basis orthogonality, but may do approximately so for simple data structures. In such cases it brings further insight as demonstrated by the latent factor interpretation qualitatively subsumed and shown in red bold face type on Figure 10. In a nutshell, and bearing in mind the fact that NoiseBiz is poorly represented in the PC1-2 factorial plane, the newly rotated factors' projection shows that:

- Many modalities play a role in the construction of varimax-PC2. We observe that `Traffic` and `NoiseConst` are two *pure* and anti-correlated factors in varimax-PC2, quasi-absent from the construction of varimax-PC1
- Except for `Traffic` and `NoiseConst`, all other factors also play a role in the construction of varimax-PC1.
- The recurring SRCs, in particular in the borough of Manhattan, about construction noise (`NoiseConst`), appears to displace or be displaced by other noise related complaints to varying degrees and by SRCs about `Traffic` nuisance (outside traffic noise). In other words where construction noise related SRCs increase, all other complaint modalities tend to decrease and reciprocally, to varying extents, except for environmental protection and sanitation SRCs (`EPsani`), and for urban infrastructure (`UrbInf`) SRCs.

Varimax-PC1:

That dimension reveals two tendencies among NYC dwellers and their ZIP code areas. Those most sensitized to noise either caused by car traffic during the day, or by residents at night. That group seems to report grievances under `NoiseTraf` and `NoiseResid` either with no correlation or anti-correlated with other SRC modalities. We call them the "**Noise protesters**".

Opposite on Figure 10 are areas, where citizens tend to report substandard urban conditions or services in a way apparently anti-correlated with the perception by others of noise pollution. We dubbed members of this group the "**Quality seekers**".

Varimax-PC2:

That dimension is consistent with NYC areas where inhabitants are primarily concerned by different form of urban pollution, such as: noise caused by construction work, urban sanitation, environmental issues as well as an insufficiently well-maintained urban infrastructure. We dub this group: "**City watch**".

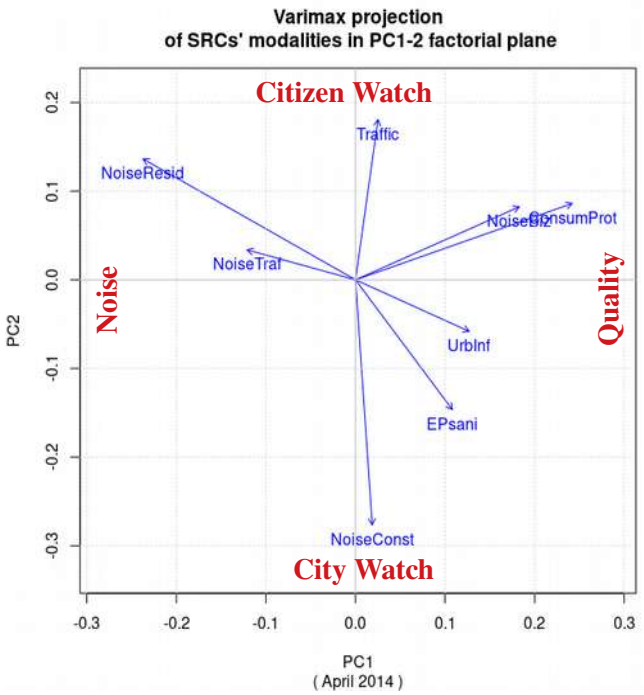


***Fig. 10**: Maximized significance of projected variables in the rotated first factorial plane PC1-2 (using the varimax method).*

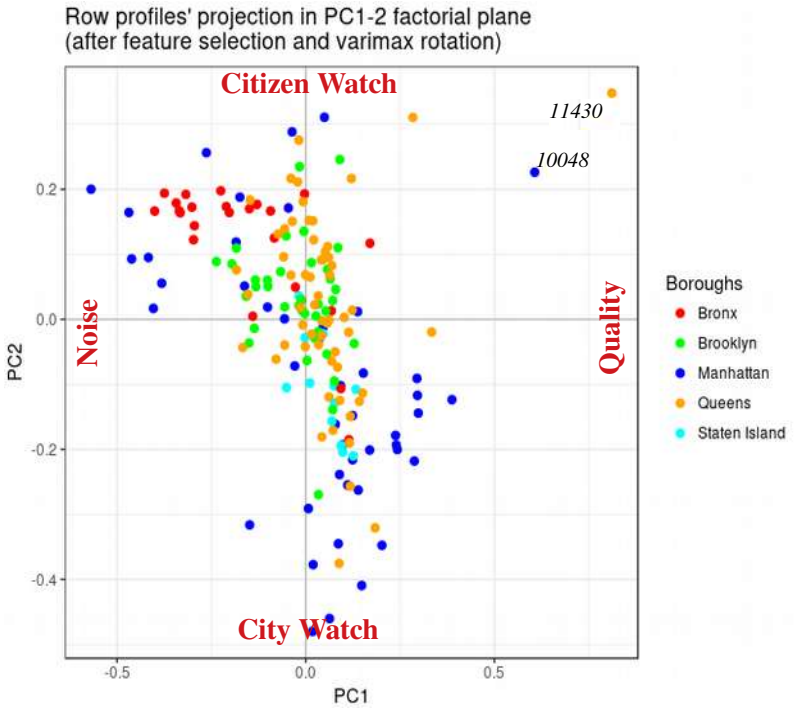


***Figure 11:** Orthogonal projections of individuals (ZIP code observations) onto the varimax-rotated loading directions, color-coded per borough and after feature selection.*

Opposite the "City watch" group, on Figure 10, are ZIP code areas, where citizens are more concerned about noise in their proximity and not caused by construction work, as well as consumer protection. Those people are also more likely to report parking violations than construction noise and are generally more sensitive to uncivil behavior as well as other nuisances directly caused by fellow urbanites. We dub this group: "**Citizen watch**".

Two figures complete the presentation of our correspondence analysis for April 2014 SRCs data. Figure 11 revisits Figure 8, the projection of scores on the first factorial plane (PC1-2), after varimax-rotation of the loadings. Figure 12 further complements that by offering a topographically situated, color-coded representation of ZIP codes' quadrants following Figure 11.

From Fig.11 it is easy to further observe that:

- 10048 (the Manhattan vanity ZIP code for the former World-Trade twin towers) and 11430 for Jamaica in Queens, resemble outliers due to very few SRCs originating in them, besides SRCs in the category "ConsumProt".

- Manhattan exhibits influences between "*City Watcher*" and a combination of "*Noise*" sensitized areas and "*Citizen Watch*"

- The Bronx is clearly dominated by residential noise related SRCs, in the graphical sector between "*Citizen Watch*" and "*Noise*".

- Brooklyn, Queens and Staten Island are dominated by varimax-rotated PC2, i.e. along latent factor axis defined by "*Citizen Watch*" and "*City Watch*".

The method used for obtain the above varimax-rotated scores, making possible the post-varimax visualization of scores (i.e. row-profiles projection along PC axes) requires a brief explanation. In varimax, loadings (i.e. eigenvectors scaled by the square roots of their respective eigenvalues) are rotated. In other words, eigenvectors obtained from the covariance matrix on scaled observations are not directly rotated. In fact, rigorously speaking, varimax rotation does not generally produce orthogonal loading vectors (even though the varimax rotation is often referred to as an orthogonal transformation) [3]. The upshot is that the orthogonal projections of individuals onto the rotated loading directions, that is the varimax rotated scores, cannot be computed in a straightforward way. To find them, one can use varimax-rotated loadings, multiplying the scaled (i.e. in our case merely centered) data by the transposed pseudo-inverse of the rotated loadings.

Figure 12: *Topographical representation of the latent factors at play in NYC's five borough area. As before in Figure 9, boroughs' ZIP code areas are drawn following the color coded legend provided in the upper-left corner of the map.*

Finally Figure 12 below illustrates the topographical representation of SRCs according to the previous latent factor analysis, i.e. after varimax rotation. Proposed latent factors (in the sense of Figure 11) found dominant for each ZIP code are shown by means of color coded solid dots. Dots' diameters are proportional to the number of total SRCs originating in the ZIP code for that period. Each latent factors' sector of dominance is defined as the positive or negative varimax-rotated PC directions ± 45°. They correspond to 90° cones, whose apices coincide with the projected cloud's centroid on the first factorial plane and whose axes of symmetry are the positive or negative rotated PC directions.

Row profiles' (i.e. individual ZIPs') projections, i.e. scores, in the varimax-rotated PC1-2 plane are color-coded according to the latent factor's cone they fall into:

- PC1+ cone	*(orchid)*	*QUALITY*
- PC2+ cone	*(green)*	*CITIZEN WATCH*
- PC1- cone	*(tan)*	*NOISE*
- PC2- cone	*(red)*	*CITY WATCH*

- As previously noted Brooklyn covers the complete range of SCRs modalities, as shown by the fact that the borough contains dots of all four colors.

- The Bronx (unsurprisingly at this point) is dominated by the 2 latent factors "*NOISE*" and "*CITIZEN WATCH*".

- Manhatttan does too to a lesser extent, but is clearly divided between down and midtown on one hand and uptown on the other hand. Down- and midtown are areas where dominant factors are "*QUALITY"* and "*CITY WATCH"* while uptown (i.e. north of Morningside Heights and Spanish Harlem) is clearly dominated by "*NOISE*" and "*CITIZEN WATCH*". This seems to reflect changes as much in residents' concerns and perception, as in the individual behaviors at the origin of SRCs.

- Queens is geographically divided between two wide areas: the west side, facing Manhattan and bordering Brooklyn and the east side facing the ocean and bordering Nassau county. The first one is characterized by the "*CITIZEN WATCH*" factor, while the second seems more focused on concerns about "*QUALITY*".

- Finally the population of Staten Island, as before, demonstrates its focus on urban conditions, as captured by the latent factor "*CITY WATCH*".

3-3. Multiple Correspondence Analysis (MCA)

As for the CA and PCA techniques, multiple correspondence analysis (MCA), when applied to nominal categorical data, aims at detecting and representing hidden structure in data. It does so by linearly mapping data as points in a low-dimensional Euclidean space. It is simultaneously applicable to a set of different categorical variables.

3-3-1. Discretization of data

To supplement our previous CA on SRCs to NYC 311 per location, we now add NYPD crime data, as crime report calls (CRCs) to 911, in the form of 3 modalities in increasing degree of gravity: violations (4,699 counts), misdemeanors (21,734 counts) and felonies (11,156 counts). Those events, recorded in April 2014, were distributed over 181 zip codes and 5 boroughs. In order to conduct MCA we discretized our multivariate contingency table so that every modality (column) is now expressed in the form of ordinal values related to 4 buckets (bins) of similar size or cardinality, i.e. with roughly the same counts of ZIP codes in each bucket.

Discrete re-encoding with buckets makes losing some information unavoidable, but is a practical way of dealing with contingency tables and to accommodate quantitative values of frequencies. It also allows us to extract 2 way contingency tables involving crime modalities and NYC boroughs. How we went from frequencies (counts) to ordinal variables is shown next for the sample consisting of NYPD's records of 21,734 misdemeanors in April 2014. Sample quartiles corresponding to the distribution of counts per ZIP code were:

```
Min.  1st Qu. Median   Mean   3rd Qu.  Max
0.0    32.0    87.0   120.1   179.0   705.0
```

Based on quartiles, the chosen bucket intervals were: < 33 – 87 – 178 – >178. Corresponding ordinal variable values are summarized in tabular form below. They may differ for non crime related variables (SRCs) where ordinal values may refer to a different count scale. This however is not detrimental to the correct overall multiple correspondence analysis.

Bin upper bound	2~3	6~16	20~33	~38	85~91	150~180	> 180
Ordinal variable value	VL	ML	M	MH	H	VH	OC
Interpretation	Very low	Medium low	Medium	Medium high	High	Very high	"Out of Control"

An exception is made for the treatment of the SRC categorical variable "HousCond". Its count spread is such (over several different time periods of interest) that rather than setting a common fixed bin scale, we just report quartile intervals for each time period.

3-3-2. Analysis of crime segmentation across NYC boroughs

The normalized crime segmentation per borough, for each crime modality is shown next, in Figure 13.

For each crime modality, Fig. 13 is interpreted in terms of the relative proportions of ZIP code areas in each borough belonging to a low, medium, high or very high crime count bucket. For instance, for the 4860 misdemeanors committed in Manhattan ZIP code areas in April 2104:

- 30% of ZIP code exhibited a medium (M – cyan) crime count, (14/46)
- 21% of ZIP codes exhibited a high (H – orange) crime count, (10/46)
- 26% of ZIP codes exhibited a very high (VH – red) crime count, (12/46)
- 21% belong to the bucket of extremely large crime counts, dubbed "out of control" (OC – dark red), (10/46)

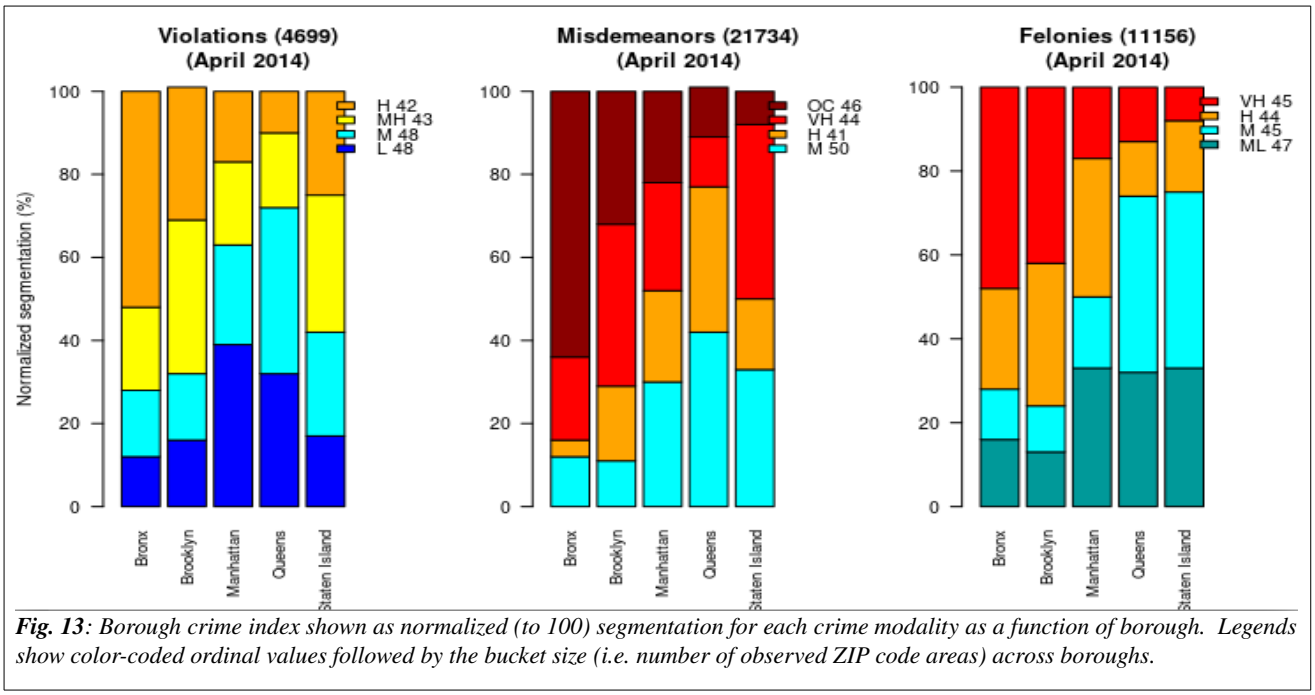


***Fig. 13**: Borough crime index shown as normalized (to 100) segmentation for each crime modality as a function of borough. Legends show color-coded ordinal values followed by the bucket size (i.e. number of observed ZIP code areas) across boroughs.*

Figure 13 does not inform us on the number of crime committed in each borough, but rather on the distribution of hot "crime spots" within each borough. The Bronx clearly shows a tendency to concentrate high crime areas, across all crime modalities, when compared to other boroughs. It is followed in that by Brooklyn, Manhattan, Queens and Staten Island, in the cases of misdemeanors and felonies.

3-3-3. MCA

We extend our previous Correspondence Analysis (CA) results to include:

- the categorical variable `Crime` (CRCs) whose three modalities are described earlier, in Section 3-3-2,
- the two quantitative variables: `medianInc` (median income) and `jlBenef` (jobless benefit).

The resulting Multiple Correspondence Analysis is based on the indicator matrix method. It specifies:

- rows "99999" (bogus ZIP code), "11430" (JFK airport, Queens), "10463" (Riverdale, the Bronx) as supplementary individuals, and
- columns `medianInc` and `jlBenef` as quantitative supplementary variables.

Figure 14 shows how all crime modalities ("`Violation`", "`Misdemeanor`" and "`Felony`") are particularly correlated with the (almost super-imposed) SRCs modalities "`HousCond`" and "`Traffic`", indicating that areas where housing conditions are poor and traffic violations reported by inhabitants are numerous also have a higher crime incidence in all three crime modalities. On the figure, the dashed gray line represents the direction of crime growth.

To note the two first PCs account in MCA for far smaller fraction of system inertia than their counterparts in CA or PCA (namely 23% vs. 73%). It is a normal consequence of the increase in dimensionality when carrying out MCA on binned (discretized) data.

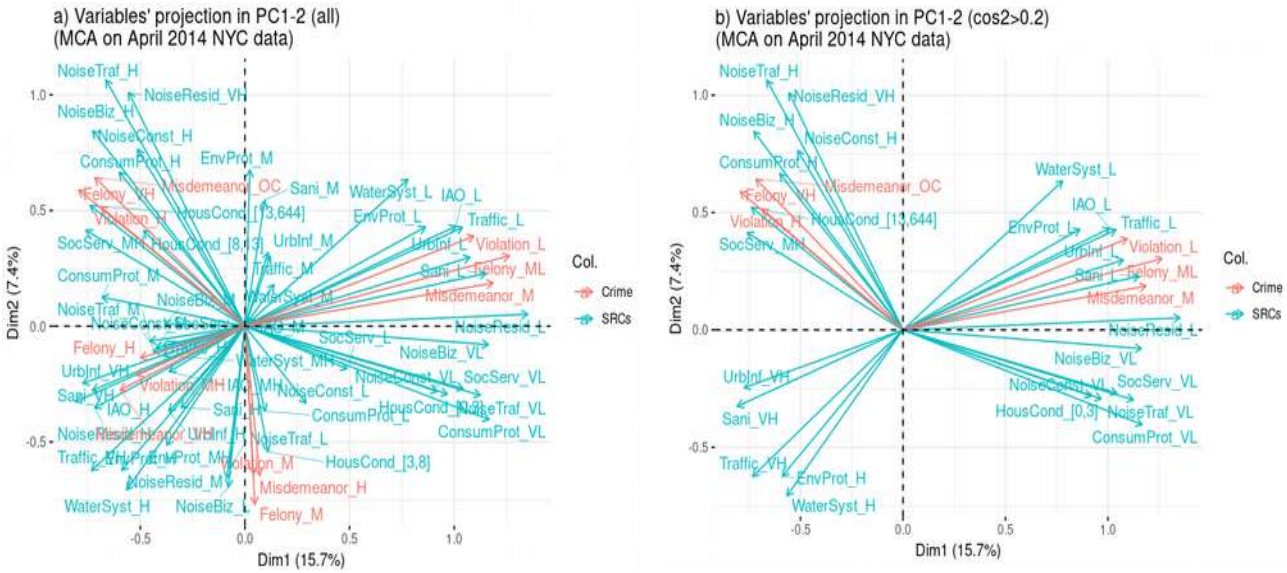


Figure 14: *MCA based representations in the 1st factorial plane (April 2014 NYC data) of variables' modalities' levels (categorical SRCs in turquoise, NYPD crime in red), for* ***a)*** *all variables and* ***b)*** *modality levels with quality of representation better than 20%.*

Fig. 15 exhibits row profiles', i.e. individuals' projection in the 1st factorial plane. The quality of representation and the contribution to the construction of PC axes are generally seen as poor compared to PCA and CA results. This is due to the inherently higher dimensionality of the MCA technique. The centroids of individuals belonging to a borough are indicated by a larger diameter colored dot. Interpretation of their relative positions in the 1st factorial plane is related to that of PC1 and PC2.

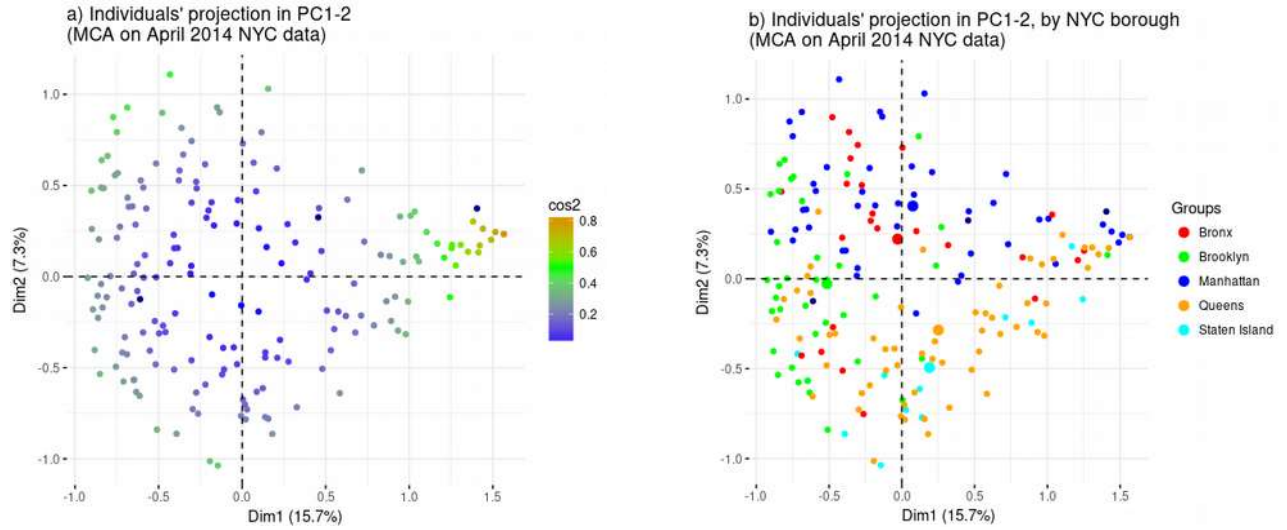


Figure 15: *PC1-2 projection for Abril 2014 of individuals (row profiles):* ***a)*** *color-coded according to quality of representation (cos2),* ***b)*** *color-coded according to NYC borough.*

Figure 16 below exhibits individual projections in the first factorial plane (PC1-2) along with crime levels per crime modality (***felony***, ***misdemeanor***, and ***violation***). Modalities' levels are represented by abbreviations as denoted before

(Section 3-2-2): low (L), medium-low (ML), medium (M), medium-high (MH), high (H), very high (VH) and out-of-control (OC) crime counts, the choice of terminology being completely arbitrary on the analyst's part. It is only meant to cover the whole scale of reported crimes counts in every category during the month of April 2014. No matter what the modality of crime is, its rate increases clock wise. Quadrant 1 contains the lowest crime rate observations and Quadrant 2 the highest.

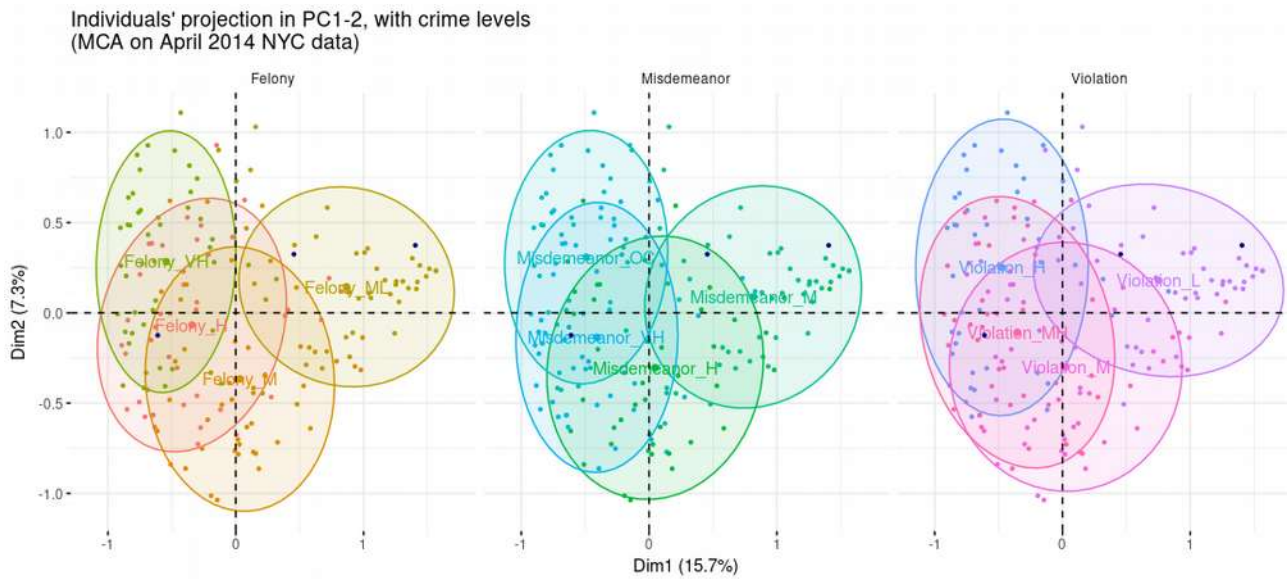


Figure 16: *PC1-2 projections (April 2014 NYC data) of individuals, color-coded according to their crime rates' modalities' levels. Ellipses are drawn for 75% confidence intervals.*

Figures 15 and 16 confirm and enrich a number of early conclusions drawn from PCA and CA.

- Manhattan consists of two parts (1st and 2nd quadrants of Fig. 15b): downtown and midtown Manhattan (dark blue dots), characterized by relatively low frequencies of SRCs (across all modalities of SRCs) and by the statistically lowest crime rate in NYC.

- Meanwhile uptown Manhattan borders the Bronx (red dots), and shares many traits with it in all crime modalities and in many SRCs. From the view point of urban planning it is a transition area between very different neighborhoods of the 5 borough metropolitan area. Going from south to north, Manhattan transitions from low to very low frequency SRCs neighborhoods to areas where complaints related to poor public housing conditions, noise (in particular but not only residential noise), traffic nuisance and reported occurrences of crimes are at their statistical highest.

- The 2nd Quadrant of Figures 15 and 16 covers mainly uptown Manhattan, South and West Bronx, Central Brooklyn as well as a few isolated ZIP codes belonging to Queens, for a total 51 ZIP codes out of 177. Those are the most violent areas in NYC in April 2014 , with:
  - a ***violation*** sum-total of 1691 representing 36% of all reported violations in the NYC area
  - a ***misdemeanors*** sum-total of 8943 representing 40% of all reported misdemeanors.
  - a ***felony*** sum-total of 4294 representing 37% of all reported felonies

Those urban areas are perceived by callers to NYC-311 as being in poor keep and the locus of uncivil or disorderly behaviors.

- 3rd and 4th quadrants are intermediate ones between highest and lowest crime rates, also between highest and lowest incidences of SRCs. The 3rd quadrant corresponds to West and East Brooklyn, a large swath of Queens plus the North-East part of the Bronx. Meanwhile the 4th quadrant reflect mainly the rest of Queens and Staten Island.

Figures 14a, 15 and 16 are combined in the form of a biplot in Figure 17, where the usual color code is used to identify the borough of each plotted individual ZIP code, crimes' modalities' levels are indicated in purple, and the rest of modalities' levels are shown in beige.

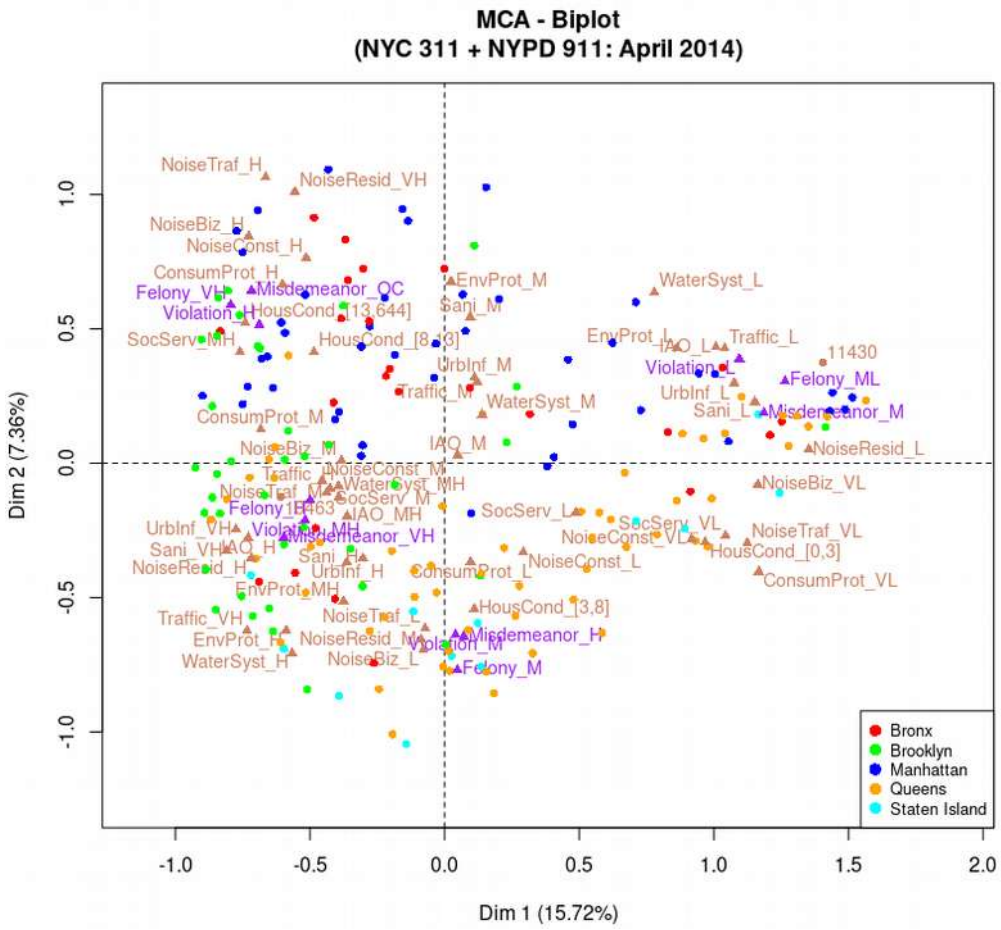


Figure 17: *MCA based biplot representations in PC1-2 and for Abril 2014 of categorical variables and individuals, color-coded according to their borough. The brown dots are the supplementary individuals:"10463" (Riverdale, Bronx) and "11430" (JFK airport in Jamaica, Queens).*

We define Figure 17's quadrant associations loosely and as follows:

- **1st Quadrant**:

- very low to medium counts of SRC's modalities "*EnvProt*", "*WaterSyst*", "*Traffic*", *UrbInf*", "*IAO*", "*Sani*" and a high count of "*ConsumProt*" SRCs.

- low incidence of reported *violations*,
- medium incidence of reported *misdemeanors*,
- medium-low incidence of *felonies*

The 1st quadrant is representative of urban pockets across the five boroughs, associated with the latent factor ***Quality***. It shows that a fairly appeased crime scene correlates well with a much reduced frequency of calls to NYC's 311.

- **2nd Quadrant**:

- medium to very high counts of SRC's modalities "*NoiseResid*", "*NoiseConst*", "*NoiseTraf*", "*NoiseBiz*", "*HousCond*", "*Traffic*", "*ConsumProt*", "*SocServ*".

- high incidence of reported *violations*,
- out-of-control reported *misdemeanors*, i.e. an incidence rate so high as to dwarf other areas in NYC.
- very high incidence of *felonies*.

The 2nd quadrant characterizes mid and uptown Manhattan, most of the Bronx and central Brooklyn, by far areas with the highest crime rate among the five boroughs, and associated with the latent factors ***Noise*** and ***City Watch***.

▪ **3rd Quadrant**:
- medium to high counts of SRC's modalities "*NoiseResid*", "*WaterSyst*", "*EnvProt*", "*ConsumProt*", "*IAO*"
- high to very high counts of SRCs' modalities "*Sani*", "*UrbInf*", "*Traffic*", "*Sani*"

- medium-high incidence of reported *violations*,
- very high incidence of reported *misdemeanors*,
- high incidence of reported *felonies*

The 3rd quadrant concerns small pockets in the Bronx and a significant part of Brooklyn as well as four Staten Island's ZIP codes and a sizable area of Queens. It appears associated with the latent factors ***City Watch*** as far as Staten Island ZIP codes are concerned,and with ***Citizen Watch*** for ZIP codes associated with either the Bronx or Brooklyn.

▪ **4th Quadrant**:
- very low to low counts of SRC's modalities "*NoiseResid*", "*HousCond*", "*NoiseConst*", "*NoiseTraf*", "*NoiseBiz*", "*ConsumProt*", "*SocServ*".

- medium incidence of reported *violations*,
- high incidence of reported *misdemeanors*,
- medium incidence of *felonies*

The 4th quadrant concerns the rest of Staten Island and most of Queens and appears to be associated with latent factors ***Noise*** and ***Citizen Watch***.

3-4. Clustering analysis

To further explore the underlying data stucture in our April 2014 NYC data set, we carry out probabilistic clustering on the row profiles of our previous MCA data matrix, using replicated *k*-means [4] partitioning. Next we deploy agglomerative Hierarchical Clustering, a well known bottom-up grouping method. Finally, we consolidate our crisp clustering results using *k*-means.

Clustering consists in grouping objects or observations in non-overlapping groups, clusters or classes, based on some criterion of proximity, similarity or likeness. Its purpose is to help understanding complex information by reducing its dimensionality. The concept we put to work is based on spherical cluster classes (multi-dimensional Euclidian proximity or distance), separable in such a way that the mean observables' value in a class converge towards the class' centroid. It ensues that clusters are expected to be of similar size, for the assignment to the nearest cluster class center to be the correct assignment.

Being a Euclidian distance based classification process, *k*-means considers variance of observations, but not covariance between observations and cluster classes thereof. Superseeding this naive approach would be possible in a number of ways [5], for instance with:

- the *Gaussian Mixture* based on the *expectation-maximization* algorithm, which maintains a probabilistic assignment to cluster classes (Bayesian soft clustering) and a multivariate normal (MVN) distribution instead of the mean;

- the *Partitioning Around k-Medoids* (PAM), a heuristic algorithm reminiscent of *k*-means but which makes use of arbitrary non-Euclidian distances such as the Manhattan (L_1) distance, the Jaquart distance, the cosine similarity, etc.

Still, despite its shortcomings, the k-means technique fulfills the objectives of our exploratory multivariate analysis.

3-4-1. Probabilistic k-means and hierarchical clustering

The standard *k*-means algorithm is a heuristic process. Replication is made necessary by the fact that (a) its result depends on initial conditions, i.e. the choice of the *k* initial centers, and (b) the algorithm does not guarantee a global optimum.

Therefore, we first deployed a probabilistic clustering analysis using twenty *k*-means replications and a number of clusters to be ascertained in the range 2~10. Individuals to be clustered are embedded in an Euclidean space defined by the factorial coordinates or "scores" of our observations, derived from the MCA results of Section 3.2.

For every experiment consisting of 20 replicas each, we calculate two ratios, and use them as criteria to ascertain the optimal number of cluster classes, allowing for 3 random starts per replication.

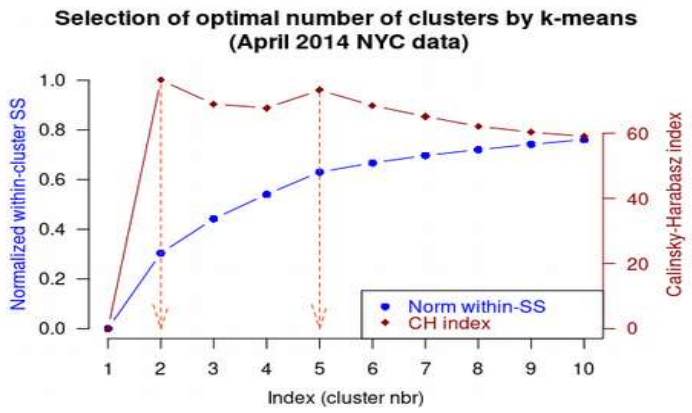


***Figure 18**: Graph of the two index criteria used to qualify our probabilistic search for the optimal number of clusters (from MCA results on April 2014 NYC data).*

- *SSB/SStot*, the between-cluster *sum of squares* or variance, denoted SS_B over the total SS", referred to as the "normalized within-cluster SS criterion", where SS represents inertia (variance) and is calculated relative to the relevant cluster centroid for each computed cluster.

- the *Calinsky-Harabasz index* consisting of the ratio of *between-cluster SS*, , denoted as before SS_B, and *within-cluster SS*, denoted SS_W, corrected by the number of clusters, k, and observations, n:

$$\frac{SS_B/(k-1)}{SS_W/(n-k)}$$

Whereas the first ratio (blue line) increases continuously as the number of trial clusters rises, we observe in Figure 18 that the CH index (red line) gives us two local optima for 2 and 5 clusters each, signalled by dashed vertical arrows pointing toward abscissae 2 and 5.

The above result is further qualified by the Cluster S*ilhouette* method, the which given k clusters and n individuals *"i"*, computes:

- *a(i)* the distance of i to all individuals of the same cluster class
- *b(i)* the lowest distance of i to all individuals of any other cluster class
- $s(i) = \dfrac{b(i) - a(i)}{max[a(i), b(i)]} \in [-1, 1]$

where *s(i) > 0*, *s(i) = 0 or s(i) < 0* when *i* is correctly allocated, close to the decision boundary or allocated to the wrong cluster respectively. Average[*s(i)*] over the whole clustered data set is a measure of clustering quality exemplified in Fig. 19:

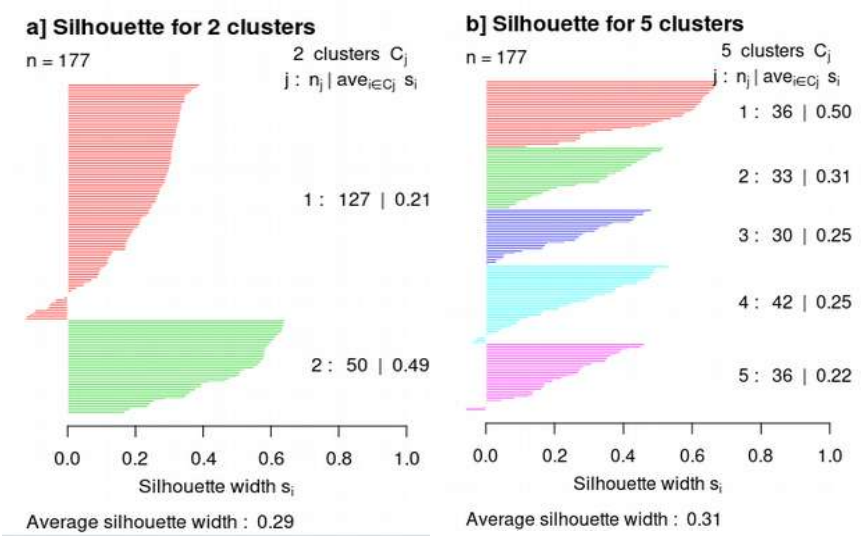


***Figure 19**: Cluster silhouette for 2 and 5 cluster classes, out of a population of 177 individual observed during April 2014 NYC (SRCs + CRCs) data.*

***a)** The 2 cluster silhouette shows an imbalance in the numbers of allocated individuals between the 2 clusters. It also sports a large number of erroneously classified individuals (ZIP codes). Its average cluster silhouette width is: 0.29.*

***b)** The 5 cluster silhouette contrasts in that its shows a well balanced distribution of observations among clusters and very few allocation errors. This contributes to the better average cluster silhouette width of 0.31.*

We further substantiate the preliminary finding of 5 cluster classes, rather than just 2 (trivial result), by representing the dendrogram built from hierarchical clustering, in Figure 20. Conceptually, considering ZIP code level agglomeration is consistent with nested hierarchies in Ward-similarity based HC. For the Ward's method, the similarity between two clusters (or two cluster-classes) is defined as the increase in the squared error that results when two clusters are merged.

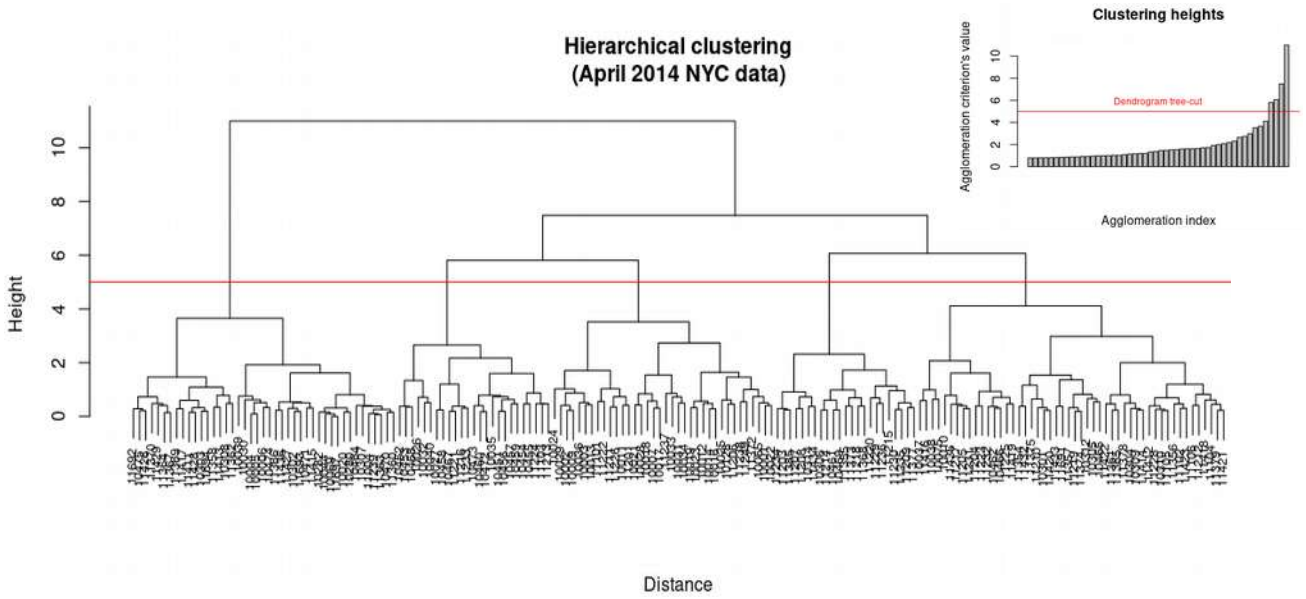


Figure 20: *HC dendrogram based on a Ward-2 distance matrix deterministically calculated from MCA scores. The most appropriate tree pruning corresponds to 5 clusters, per the horizontal red line. The inset (top-right) offers another graphical view of the agglomeration's criterion's values for each one of the 50 last merge-operations.*

Figure 21 hereafter exhibits a PC1-2 factorial plane projection of observed ZIP codes according to cluster and to borough. Appendix D contains the same plot with fully labelled observations.

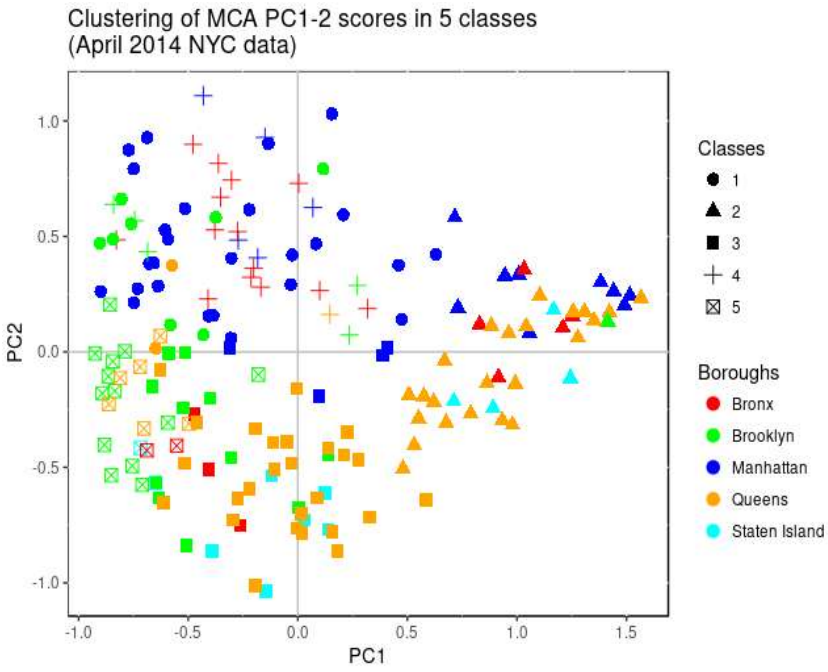


Figure 21*: Grap of scores in the first factorial plane, where each ZIP code area is shown to belong to a specific cluster class, identified by a distinct symbol. Observations are color-coded following Figs. 11, 15b, 17 and others.*

Boroughs are not uniformly clustered.

The Bronx (red symbols) is shown to belong to 3 distinct clusters (3, 4 and 5).

Brooklyn (green symbols) ZIPs are primarily distributed in clusters 1,3 and 5.

Manhattan ZIP code areas (dark blue symbols) are mainly seen in cluster 1 and 2 with very few ZIP observations classified in 3 and 4.

Queens (orange symbols) appears in clusters 2,3 and 5.

Staten Island ZIP codes (cyan symbols) are divided between cluster 2 and 3.

Before completing this analytical sequence for the period of April 2014, we show a comparative summary in terms of Inertia Explanatory Power (IEP), as obtained from CA/PCA without crime data on one hand and from MCA with crime data on the other hand.

The inclusion of crime statistics brings about a shift in IEP per borough, as evidenced in Table 7.

- The Bronx and Staten Island conserve their significance in terms of IEP.
- Meanwhile Manhattan loses its statistically prominence and goes from 46% to 29% of EIP over the same period.
- Brooklyn and Queens in turn gains in significance and go from 16 to 24% and from 17 to 28% respectively.

If one considers service request calls (SRCs) and crime report as "statistical events", then the petulance of Manhattan dwellers, when it comes to reporting urban nuisance by calling NYC 311, appears "statistically diluted" by the crime rates in the two neighboring counties of Queens and Brooklyn.

At this stage hierarchical clustering as conducted by us also shows (viz. Fig. 21) that at least 3 dimensions are at play. This is made obvious by the fact that the 5 detected clusters appear (to a large extent) superimposed in the first factorial plane, but not so when looked at in 3 or more dimensions.

Figure 22 offers a basic 3D perspective for the five cluster classes clearly separated in 3D space.

Borough	*number of ZIP codes*	*IEP all_dim PCA (%)*	*IEP 5_dim MCA (%)*
Bronx	*24*	*13.9*	*13*
Brooklyn	*38*	*16.2*	*23.4*
Manhattan	*46*	*46.2*	*29.1*
Queens	*59*	*17.2*	*28.1*
Staten Isl.	*12*	*6.0*	*6.5*

***Table 7**: Inertia explanatory power by ZIP codes, grouped by borough, computed over all dimensions (col. 3 - with PCA, SRCs w/o crime data) and over the 5 most significant significant dimensions (col. 4Th – with MCA, SRCs w/ crime data).*

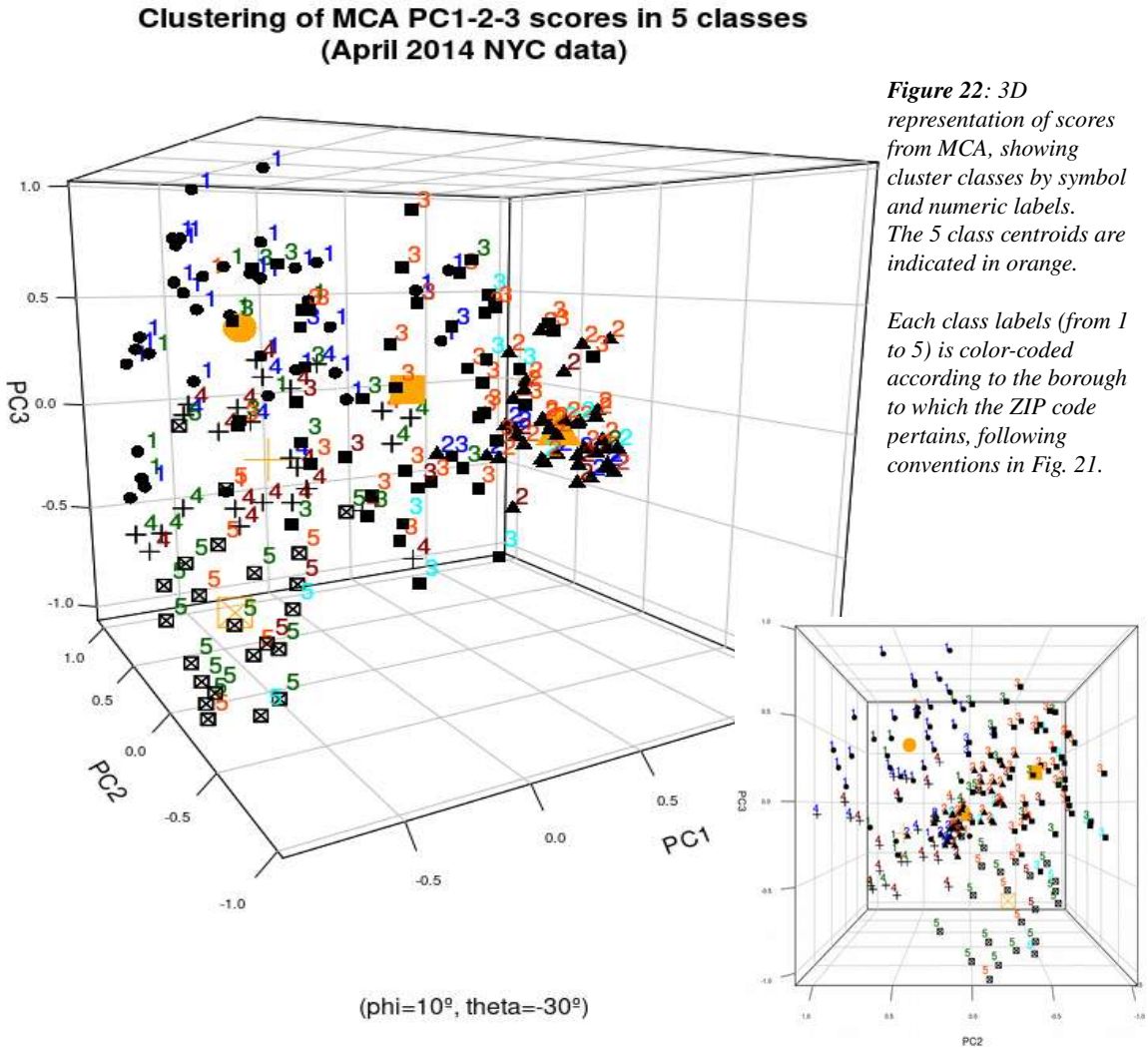


***Figure 22**: 3D representation of scores from MCA, showing cluster classes by symbol and numeric labels. The 5 class centroids are indicated in orange.*

Each class labels (from 1 to 5) is color-coded according to the borough to which the ZIP code pertains, following conventions in Fig. 21.

The inset (bottom right) in Figure 22 above offers a perspective in the PC1-3 plane (i.e. phi=0º and theta=90º) with the 5 orange colored cluster class centroids. It reveals the importance of the third factorial dimension (PC3) for out of 1st factorial plane structure, and exhibits clearer cluster separation in 3D space.

3-4-2. Clustering with k-means consolidation

As pointed out before, we first deployed a probabilistic approach to ascertain the optimal number of clusters describing our data set. The class centroids resulting from hierarchical clustering are then used as seeds to conduct a new k-means computational optimization of classes. The so-called "clustering consolidation" technique permits overcoming to some extent the curse of "merges being final" in Ward2-based Hierarchical Clustering. A qualitative explanation follows.

Although the minimum sum-of-squares criterion is used in HC with Ward2, ensuring that no merge occurs if the system's resultant Sum of Square is not minimized, it is in fact possible to merge observations (or groups thereof) even though the merged "points" may be closer to another cluster's centroids than to the centroid of its current cluster. In that sense the sequence of successive merges in HC is path-dependent and therefore not optimal. By using centroids so obtained to conduct a new *k*-means optimization, a reshuffling of observations occurs about them, while the same centroids continuously updates their positions until convergence. This yields improved clustering results and remedies the difficulty inherent in performing HC with Ward2.

Figure 23 exhibits the modified scatter plot of MCA scores in the 1st factorial plane, *after consolidation*, using both color codes and character symbols of Figure 21.

As expected class members are somewhat redistributed. Most notably classes 1 (purple) and 3 (tan) appear to lose membership, while the numbers of others either remain stable or grow. By this consolidation method, the quality index is: $I_b = \frac{SS_W}{SS_W + SS_B}$ It can only improve. In the present case, its value increases from 60.3 to 62.8.

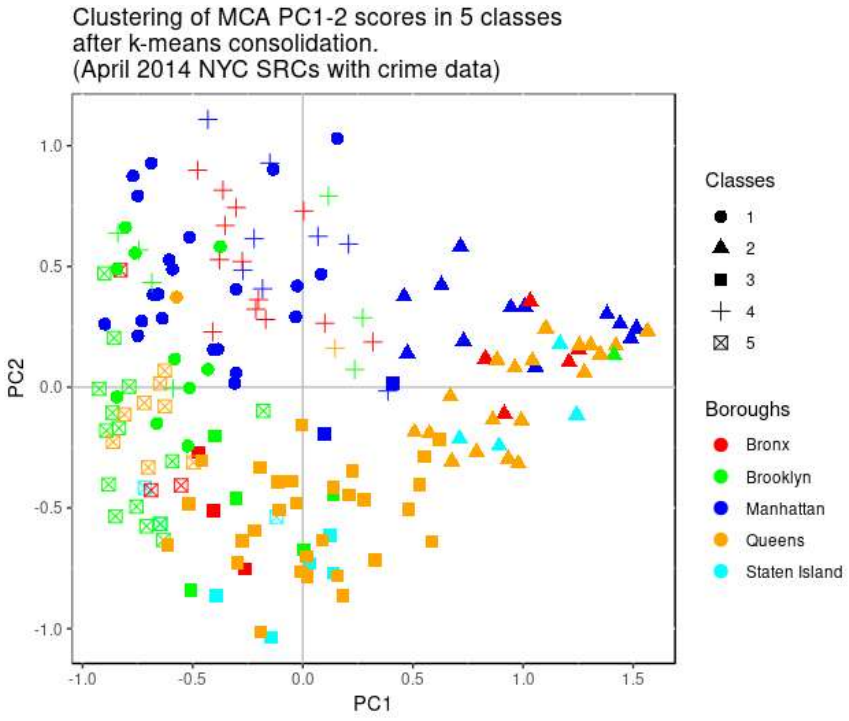


***Figure 23**: Graph of scores in the first factorial plane after k-means consolidation, where each ZIP code area is shown to belong to a specific cluster class, identified by a distinct symbol. Observations are color-coded following Fig. 21.*

Each class' member ZIP codes are topologically mapped in Figure 23 after consolidation. As before (in Figure 21) boroughs are not uniformly clustered. Rather:

Bronx (red symbols) ZIP codes are distributed among 4 distinct classes (4, 2, 5 and 3), class 4 being prominent.
Brooklyn (green symbols) ZIP codes are distributed among the 5 classes, class 5 being prominent.
Manhattan (blue symbols) ZIP codes are mainly seen in cluster classes 1 ,2, 4 and 3, in decreasing order of importance.
Queens (orange symbols) ZIP codes appears in clusters classes 3, 2 and 5, in decreasing order of importance.
Staten Island (cyan symbols) ZIP codes are divided between cluster 3, 2 and 5, in decreasing order of importance.

Additionally each cluster class' size and IEP is listed in Table 8 below before and after consolidation. Cluster classes' colors correspond to numeric labels and to character shapes (but not to colors) in Figure 22 and to cluster class hues (bottom right legend) in Figure 23.

As the main effect of *k*-means consolidation, we see that cluster class 3 (square symbols in Figures 21 and 23 and tan-colored dots in Figure 24) loses a lot of inertia representativeness: it goes from 21.8 to 14.8% in favor of classes 4 and 5, which increase each to 18.2%.

Figure 24 is reminiscent of Fig 9, where ZIP code areas characterized by noise (in particular residential), traffic nuisance, a high crime rate and poor housing conditions (now signalled by dark red dots) seem better circumscribed than before as Cluster class 4.

	Before consolidation		***After consolidation***	
Cluster class	***number of ZIP codes in class***	***IEP 5_dim (%)***	***number of ZIP codes in class***	***IEP 5_dim (%)***
1 (●)	*36*	*21*	*33*	*20.3*
2 (▲)	*43*	*28.6*	*42*	*28.6*
3 (■)	*50*	*21.8*	*44*	*14.8*
4 (+)	*25*	*13.5*	*29*	*18.2*
5 (☒)	*23*	*15.1*	*29*	*18.2*

Table 8*: Inertia explanatory power by ZIP codes observation scores (SRCs w/ crime data), grouped by cluster class, computed over the 5 most significant significant dimensions , before and after k-means consolidation.*

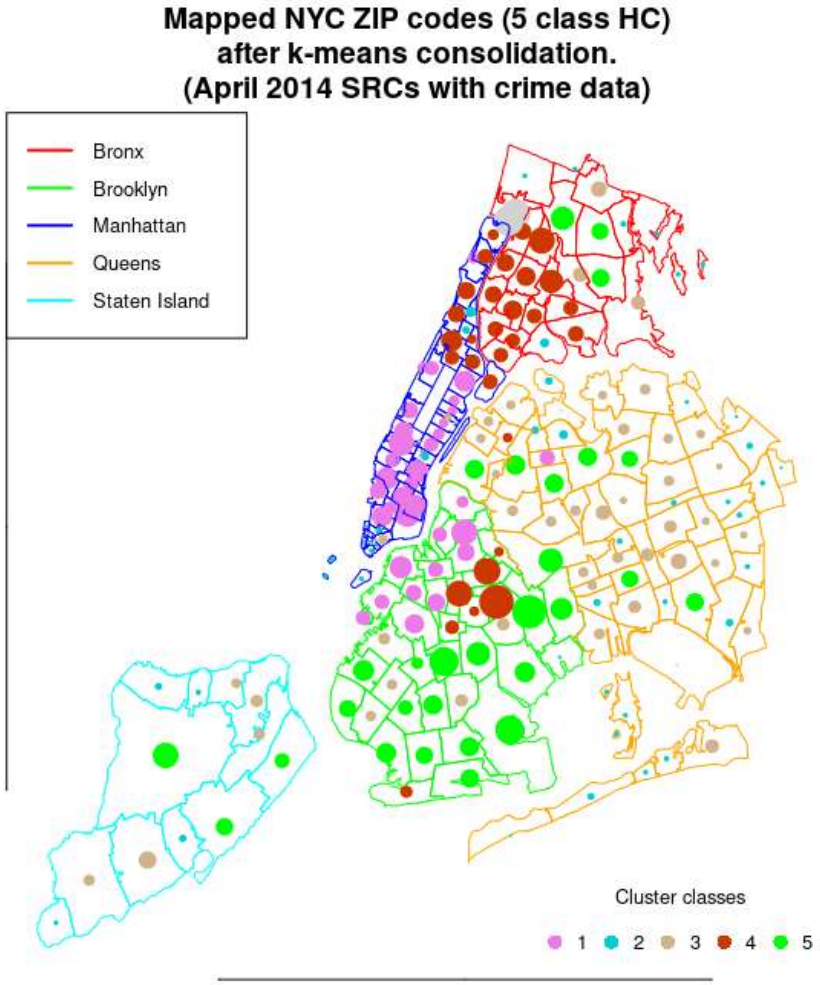


Figure 24*: Topological mapping of the 5 class cluster obtained from Hierarchical Clustering (HC),* ***after k-means consolidation****. Gray colored dots are either outliers or ZIP codes areas otherwise not included in the analysis.*

■ Recalling that dot diameters are directly proportional to the frequency counts in a given ZIP code area, we see that adding crime statitics to SRCs has a balancing effect on dot size across cluster classes.

■ We nonetheless distinguish class 2 (cyan colored dots as the group of ZIP codes with the lowest incidence of recorded events and class 3 as the second lowest.

■ As previously noted, Manhattan now appears statistically comparable to other boroughs, if not in terms of the nature of recorded events, then at least in terms of event frequency.

The same topological representation, obtained from Hierarchical Clustering **without** *k*-means consolidation, is provided in Appendix E. It differs from Figure 24 in significant aspects.

To go beyond the mere visual inspection of the distribution cluster classes in the NYC geography, we need to identify quantitatively which factors, and among them which categorical variable modality, significantly contribute to the construction of clusters. For that we execute tests of independence between cluster classes (5 classes mean 4 DoFs) and each categorical variable (4 modalities means 3 DoFs). Results are summarized in Table 9, where we see that the null hypothesis (H0: "there is no relationship between the 2 tested categorical variables.") is rejected for each variable but to various degrees.

The variables most related to the formation of classes are, in decreasing order of significance:
ConsumProt, Felony, NoiseConst, NoiseResid, NoiseBiz, Misdemeanor, Sani
covering 3 orders of magnitude in values of p-values.

Inversely the least related one are:
WaterSyst, SocServ, EnvProt, IAO
as was already intimated during the PCA, CA analyses.

Figures 24 and 25 below provide elements to interpret the latent class semantics. Figure 24 in particular show how certain variable modalities may play a role, i.e.:
- be significantly over-represented (in blue) simultaneously in up to 3 classes,
- be significantly under-represented (in red) simultaneously in up to 2 classes.

This induces a certain complexity of interpretation, better unraveled by Figures 25 on the basis of which, we propose a summary view of each class profile.

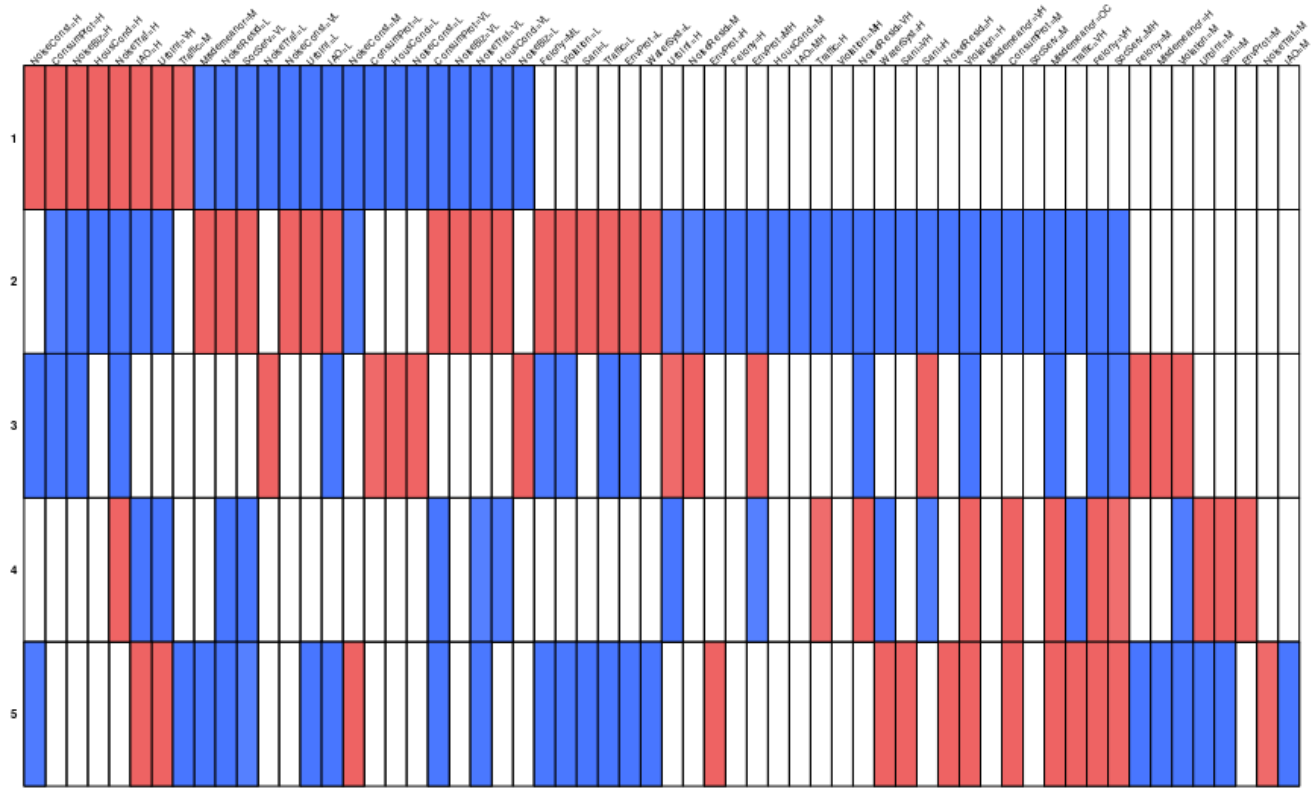


Figure 24*: Matrix view of over-represented (blue) and under-represented (red) modalities in each class (rows 1 to 5).*

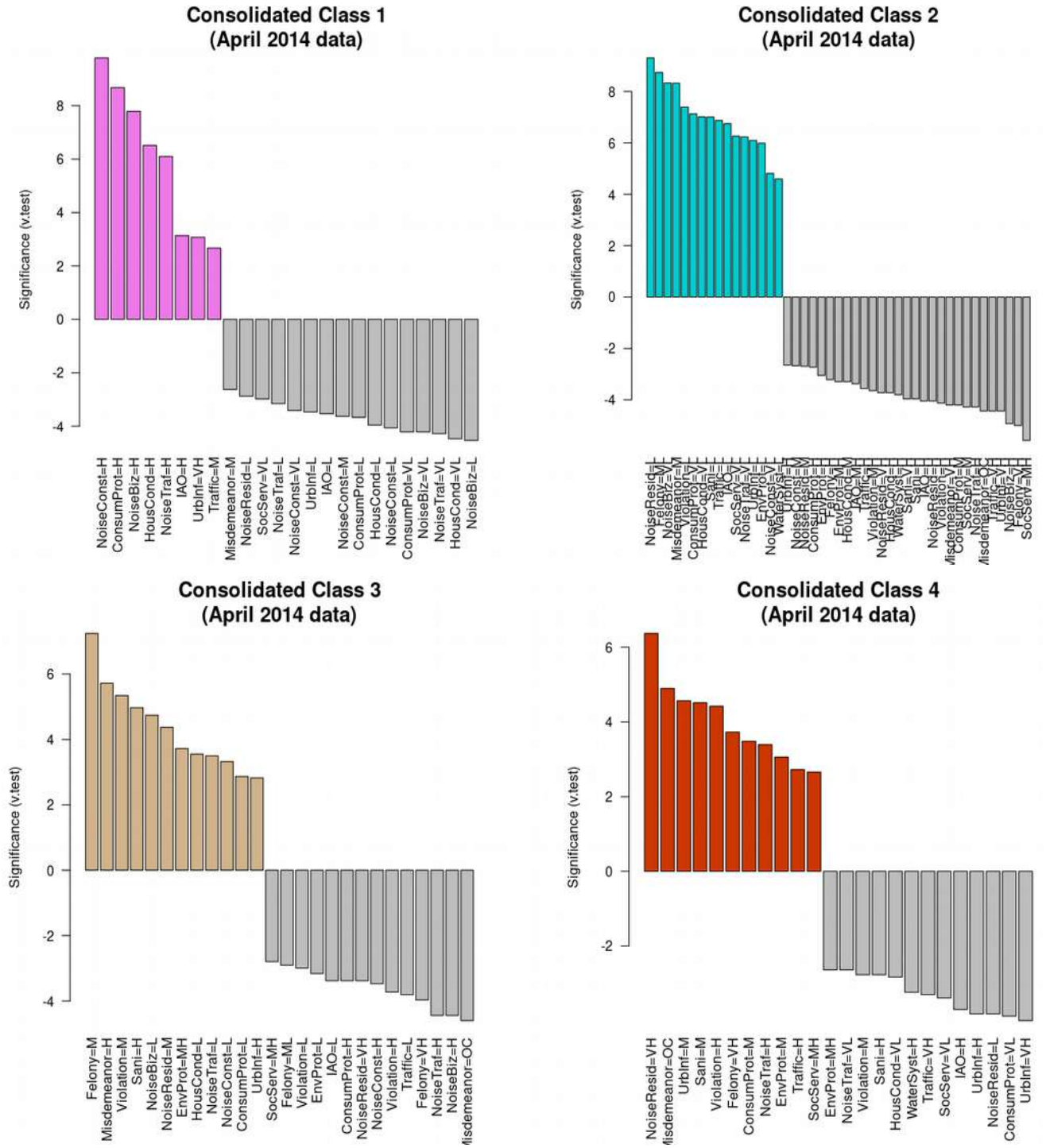


***Figure 25**:Significance of modalities for classes 1,2 3 and 4 (NYC SRCs and crime data). Over-represented modalities are color-coded per class following the convention adopted for Table 8, while under-represented modalities are shown in gray.*

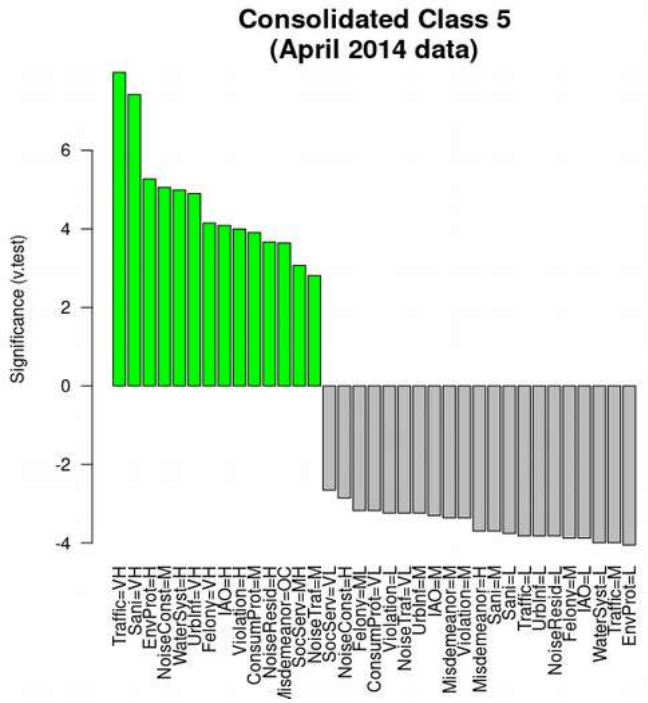


***Figure 25 (continued)**:Significance of modalities for class 5 (NYC SRCs and crime data). Over-represented modalities are color-code following Table 8, while under-represented modalities are shown in gray.*

Class 1:
High (H) incidence of SRCs: "*NoiseConst*", "*ConsumProt*", "*NoiseBiz*", "*HousCond*" and "*NoiseTraf*".
Crime related modalities do not play a significant role in the construction of that class.

Class 2:
Low (L) to very low (VL) incidence of SRCs, normally associated with dense urban areas, traffic, public housing, sustained street-level commercial activity ("*NoiseResid*=L", "*NoiseBiz*=VL", "*ConsumProt*=VL", "*HousCond*=VL", "*Sani*=L", "*Traffic*=L", "*NoiseTraf*=VL", "*UrbInf*=L", "*NoiseConst*=VL")
Crime related modalities are significant and show a moderate to low crime rate: "*Felony*=ML", "*Misdemeanor*=M", "*Violation*=L".

Class 3:
Salient SRC modalities show medium to high concern for the condition of public places and urban infrastructure ("*Sani*=H", "*EnvProt*=MH", "*UrbInf*=H"),
- moderate residential noise related calls: ("*NoiseResid*=M"),
- low levels of complaint about factors normally associated with public housing, commercial areas, and construction work: ("*NoiseBiz*=L", "*HousCond*=L", "*NoiseTraf*=L", "*NoiseConst*=L", "*ConsumProt*=L")

Crime related modalities are significant with slightly higher level of incidence than for Class 2: "*Felony*=M", "*Misdemeanor*=H" "*Violation*=M".

Class 4:
Medium (M) to very high (VH) incidence of SRCs, usually associated with high population densities, public housing infrastructure and lower wealth: "*NoiseResid*=VH", "*UrbInf*=M", "*Sani*=M", "*ConsumProt*=M" "*NoiseTraf*=H", "*EnvProt*=M", "*Traffic*=H", "*SocServ*=MH".
Crime related modalities play a primordial role, with very high crime levels: "*Misdemeanor*=OC", "*Violation*=H", "*Felony*=VH".

Class 5:
High incidence of SRCs: "*Traffic*=VH", "*Sani*=VH", "*EnvProt*=H", "*NoiseConst*=M", "*UrbInf*=VH", "*ConsumProt*=M", "*NoiseResid*=H", "*SocServ*=MH", "*NoiseTraf*=M"
Crime related modalities play a significant role in the construction of that class ("*Felony*=VH", "*Violation*=H", "*Misdemeanor*=OC"), at a level identical to that of class 4.

The main differentiating factors of Class 5, when compared to Class 4, are:
"*Sani*=VH", "*EnvProt*=H" and "*UrbInf*=VH" as well as "*NoiseConst*=M", the latter being an SRC variable which plays no significant role in the construction of Class 4.

4. Temporal evolution of NYC's urban semantics

Studying the evolution in time of urban characteristics is tantamount to comparing results for a given time window to results for a different time window, ans so forth for each period of interest. In our case we are interested in the comparison of April

2010, April 2014 and April 2018. The corresponding analytical results for the two periods ***April 2010*** and ***April 2018*** are available in Appendices F and G respectively.

A basic difficulty lies in quantitatively comparing analytical results derived from different data sets, as each data set was independently subjected to the same analysis. However careful we are in decomposing measurement space in a) a principal component (PC) space, which contain common-cause variability, and b) a residual space that contains system noise, preserving intangibles between each data set (i.e. ensuring that PCs are either the same or "close" for each data sets) is not always possible. In other words data sets generally lead to distinct orthonormal bases, as obtained from PCA, CA or MCA. Similarly cluster classes obtained from k-means and hierarchical clustering (in our case), should vary in number as well as in nature, i.e. in the factorial and individual contribution to their construction.

Broadly speaking two approaches are possible to tackle that difficulty:

1) One may choose one period of reference and consider that all observations outside that period are supplementary individuals. As such they do not contribute to the construction of the R^p factorial space and to that of principal directions. The advantage is that we need not pay attention to how far apart centroids for each data sets are from one another, since only one reference data set is included as the set of active variables with one active centroid as a result. The clear disadvantage of that method is that it fails to detect PCs different from those of the reference data set, but characteristic of the other, now illustrative, data sets. In such a case the perceived time evolution pattern between different observations periods (different data sets) will not take into account those *undetected* PCs. This generally entails a distortion in the obtained evolution pattern(s). How to detect its severity or bound its estimation is outside the scope of this work.

2) One may choose instead to analyze the three periods as one data set. In that case all observations (apart from outliers) are considered as active individuals and categorical variables are required to be the same for all data sets. In other words, data sets may be joined vertically, i.e. row-wise. In such a case a common subspace, the new R^p factorial space is created, and is expected to be different (in the general case) from that of each data set analyzed independently. The system inertia is now composed of that of each data set, plus that between data sets, due to the separation between each data set's specific centroid and the specific contribution of each data set to the construction of the global intangibles. In the special case of low or zero between-data set inertia and very close or similar intangibles among the three data sets, the two approaches must give very similar or equivalent results.

In this section, we opt for the second approach and visualize each data set's specific centroids graphically as well as their specific IEP. We then proceed to compute and plot topographical heat maps across NYC's 5 boroughs to represent change hot-spots. Here "change" means how much change in urban characteristics (of semantics) was captured for each observed ZIP code area across time.

4-1. MCA for the all-encompassing data set

As before for PCA and CA and MCA, the reader should be careful in interpreting proximity on factorial plane projections. A minimum set of simple qualitative rules applies:

1/ Proximity between two projected points is lent more and more credence as those points are situated farther from the origin of the factorial plane.

2/ In biplots the interpretation of proximity between a categorical modality level and an observation is always hazardous. On such graphs, proximity may be usefully interpreted only between observations or between categorical modalities, i.e. between points of the same set of factors.

3/ When the projections of two row profiles are close together, they tend to be similarly represented by categorical variables, in particular in the matter of levels.

4/ When the projections of *different* categorical modalities levels are close together, then such levels tend to appear together (in association) in the observations.

5/ When the projections of different levels of the *same* categorical modality are close together, and because observations may not possess different levels of the same categorical modality at once, we conclude that "*groups of observations associated with those distinct levels are themselves similar*" – cit. Abdi et al (2007) [6].

MCA encodes data by artificially creating additional dimensions. In doing so, modality feature levels are coded according to several columns in the complete disjunctive table (aka, the indicator matrix used in our implementation). Again quoting Abdi et al. (2007) [6] it follows that "*the inertia (i.e. variance) of the solution space is artificially inflated"* and the inertia explanatory power of each principal component correspondingly underestimated. The same reference mentions cursorily 2 methods for the correction of MCA eigenvalues. Both have relatively straightforward implementations. The simplest correction is customary and was implemented in this work, although this fact is not reflected by values of inertia explanatory power (IEP) shown as percentile values on the PC1 and PC2 axes of Figures 26-a to 26-f above, as well as of Figures 27-a,b below. The corrected renormalized percentile values of inertia represented by PC1 and PC2, according to Benzécri [7], are: 24.3% and 10.0%, to replace the non-corrected values of 15.8% and 6.5% respectively.

Graphical results exhibited hereafter in Figure 26-a, to Figure 26-f are a partial factorial representation in terms of contributions to the system's inertia. Bearing in mind that the first factorial plane (PC1-2) represents less than 35% of the total system's inertia, Figures 26-a to 26-f provide trajectories for observations from 2010 to 2014 and on, to 2018 for ZIP code areas in each of the five boroughs. Black triangles shown on those figures are the projections of individuals' centroids for each data set. They are such that their vector-sum coincides with the origin of the representation's factorial axes for the global data set's centered data, i.e. with point 0^p in p= dimensions.

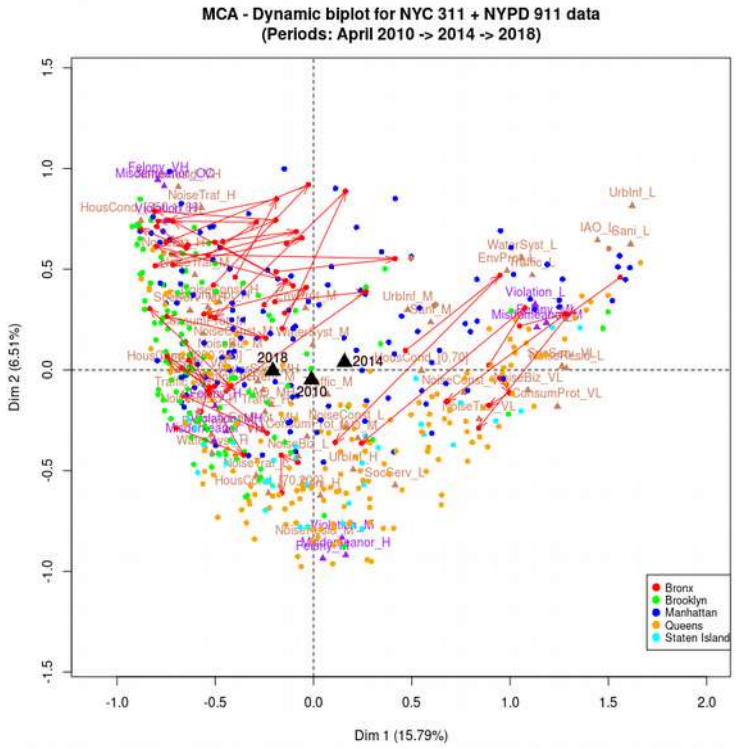


Figure 26-a*: First factorial plane (PC1-2) biplot showing the temporal evolution of ZIP code NYC 311 and NYPD 911 calls in the Bronx between April 2010, April 2014 and April 2018.*
Note*: Percentile values of IEP shown on PC1 and PC2 axis labels do not reflect the customary correction implemented for trivial inertia in MCA. Graph appearance remains unaffected.*

Urban semantics show a global move to the right of the first factorial plane, i.e. toward fewer crime report calls (CRCs to 911) and SRCs (calls to 311), between April 2010 and April 2014, followed by a distinct reversing of that trend between April 2014 and April 2018. From the point of view of citizen's perception, and barring empirical bias during data acquisition (which we are ill-equipped to test), the shifts indicate an improvement during the first 4-year period and a worsening during the second. In our context "*improvement*" and "*worsening*" are used to mean that the number of reports logged by NYC 311 and or NYPD 911 decreases (respectively increases) during the time interval no matter what the exact modalities for those varying counts are.

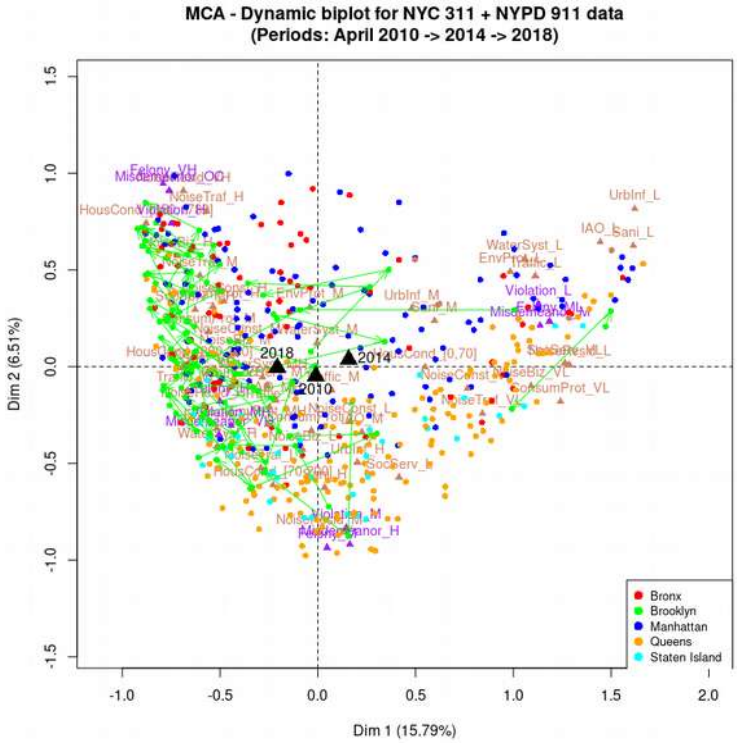


***Figure 26-b**: As for Figure 26a but for Brooklyn.*

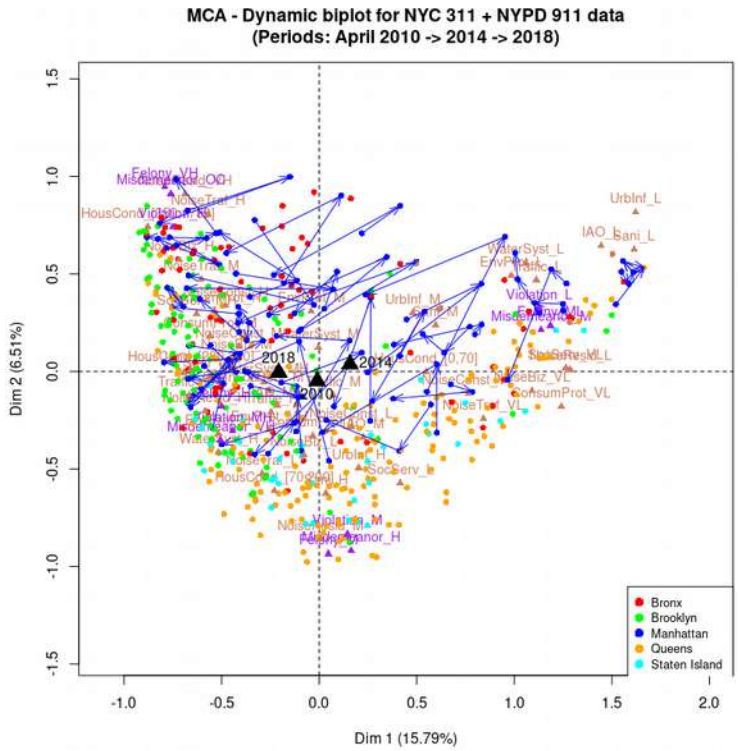


***Figure 26-c**: As for Figure 26a but for Manhattan.*

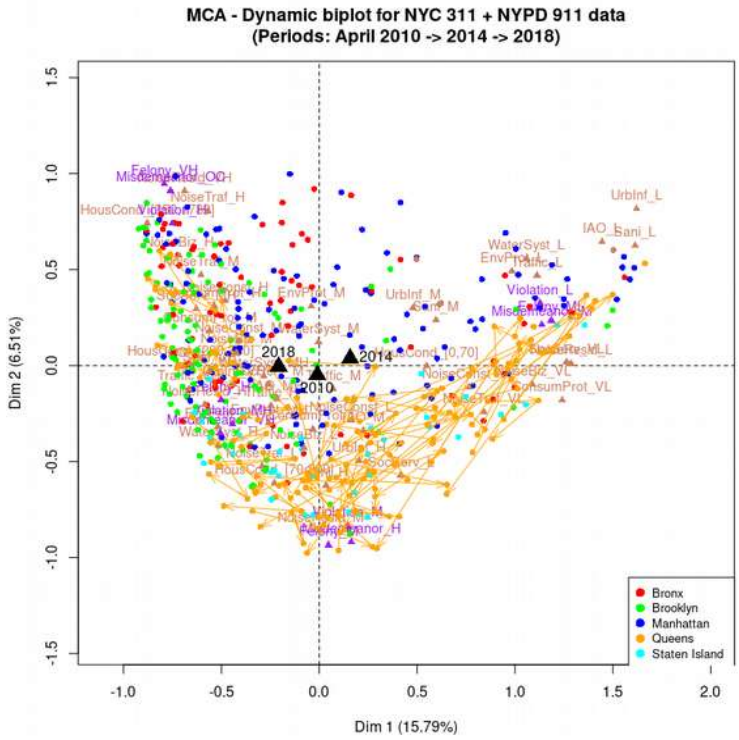


***Figure 26-d**: As for Figure 26a but for Queens.*

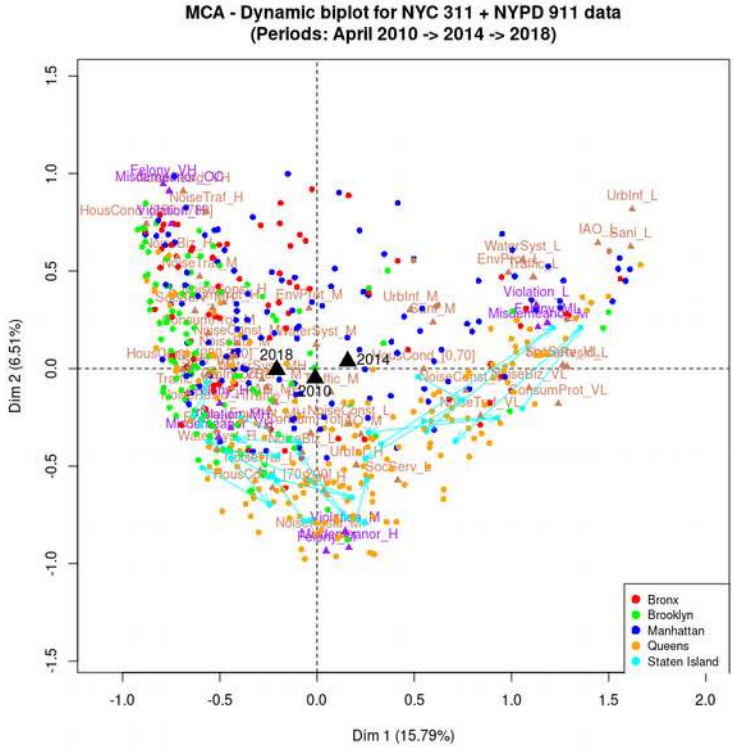


***Figure 26-e**: As for Figure 26a but for Staten Island.*

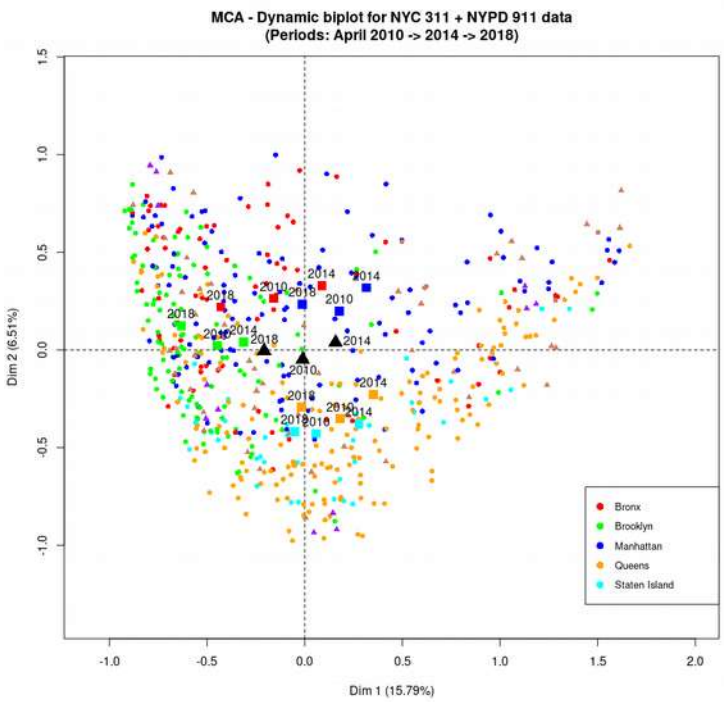


Figure 26-f*: As before, but instead of representing the temporal evolution of individual ZIP codes, individuals' centroids are represented per borough, as color-coded squares labeled with the year of the corresponding data set.*

4-2. Pertinence of our MCA approach: the all-encompassing data set

We know that the total variance of the global data set in MCA is also its total inertia. It may be computed as the chi-square statistics divided by the grand sum total for the rectangular contingency table (with positive integer valued entries). In fact, breaking the complete contingency table into smaller parts of interest is commonly tested for statistical significance in this manner, using permutation testing, aka resampling.

For our purpose it is enough to remember that *trivial inertia* in MCA is associated to within-category modality variance (i.e. to each diagonal block of the corresponding Burt matrix). Within-category modality variance computation artificially introduces added inertia, equal to the studied categories' ranks. However this inertia carries *zero additional information*.

Each data set's IEP is summarized below for the maximum dimensionality obtained from Multiple Correspondence Analysis (MCA), that is for 48 dimensions (before applying the Benzécri's correction) accounting for 100% of the information contained in the system:

Data set: April 2010
Number of observed individuals (ZIP codes): 181
Data set's IEP contribution: 32.81%

Data set: April 2014
Number of observed individuals (ZIP codes): 178
Data set's IEP contribution : 35.42%

Data set: April 2018
Number of observed individuals (ZIP codes): 183
Data set's IEP contribution: 31.63%

Adding each data set's rows' IEP yields close to 99.9% of total inertia, which is satisfactory.

We now need to ascertain the relative proportion of between-data set inertia, $\beth_B^{(tot)}$, and within-data set inertia, $\beth_W^{(tot)}$.

For the between-data set inertia: $\beth_B^{tot} = \sum_{i \in \{2010,2014,2018\}} \beth_B^{(i)}$

computed by means of the MCA coordinates, as the sum of distances of the data sets' centroids to the origin of the factorial space, divided by the number of observations in each data set, based on the number of significant dimensions.

For the within-data set inertia: $\beth_W^{tot} = \sum_{i \in \{2010,2014,2018\}} \beth_W^{(i)}$

equal to the sum of χ^2 statistics, computed for each data set's rectangular contingency table, divided by the sum total of each table's counts.

Table 9 below summarizes results:

	April 2010	**April 2014**	**April 2018**	Row-wise sum totals
Between data set inertia, $\beth_B$	0.0012	0.0020	0.0016	0.0048
Within data set inertia, $\beth_W$	0.3414	0.5319	0.4290	1.3023

Table 9*: Between and within data set inertia values for the three periods of interest April 2010, 2014 and 2018.*

We may neglect the components of inertia between data sets as their overall contributions is negligible.

We retain our method of projection based on MCA components obtained from the all-encompassing data set. In the context of our data, this approach is valid for temporal evolution representations.

A simple variables' projection in the first factorial plane may help shed some additional light on the direction of growth of each categorical variables with respect to intangibles. As before (in Fig. 14) Figures 27 below show SRC modalities in turquoise and crime modalities in red.

Previous described trends in modalities' relative positions are confirmed. In keeping with more frequent crime report in April 2010 and April 2018 with respect to April 2014, crime categorical variables' modalities' contributions are greater for 2010 and 2018 (and by extension for the global data set too), than for April 2014.

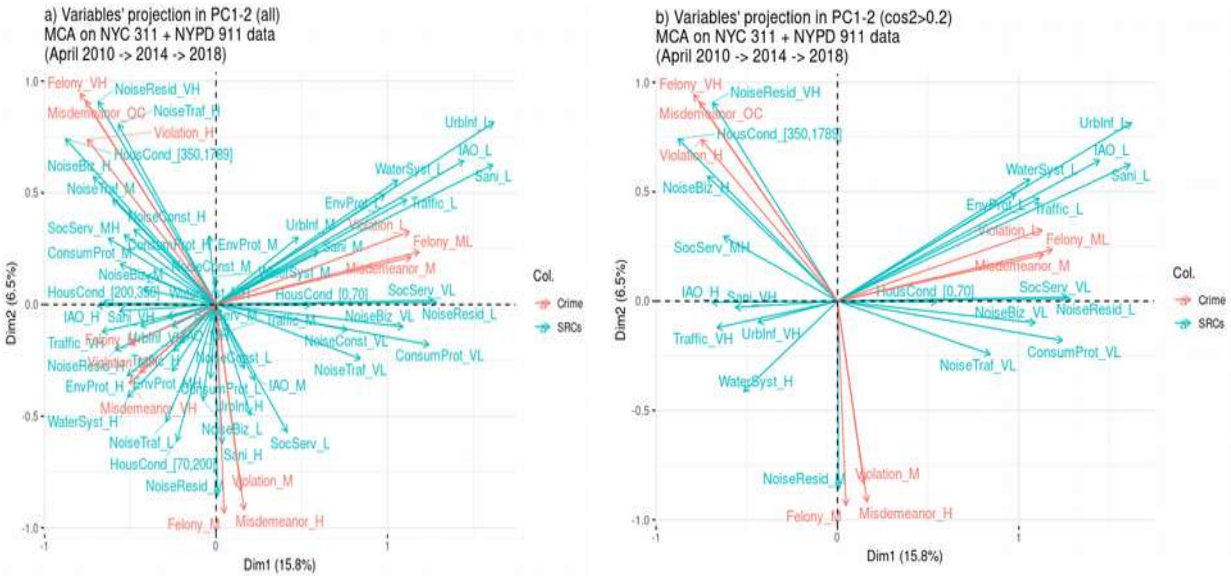


Figure 27: *MCA based representations in the 1st factorial plane of the global data set's variables' modalities' levels (NYC 311 SRCs in turquoise, NYPD 911 CRCs in red), for* ***a)*** *all variables and* ***b)*** *modality levels with quality of representation greater than 20%.*

4-3. Representation of change: topographical heat-maps

Heat maps construction is based on the pair wise Mahalanobis distance between the MCA coordinates of each ZIP code areas between two consecutive time periods. The Mahalanobis distance allows us to take into account varying correlation effects between modalities, when building a proximity metric.

As computed the Mahalanobis distance is always positive and thus cannot reflect the direction of change. When squared (as in Figures 28) , it is a modified L2 metric only representing the *intensity* of change. Intensity is indicated by the thermometer on the right hand side of each topographical map. The thermometer has a fixed scale for all 3 figures. The blue color corresponds to very small or no change, whereas change intensity increases as the color goes turns green, yellow, and finally red.

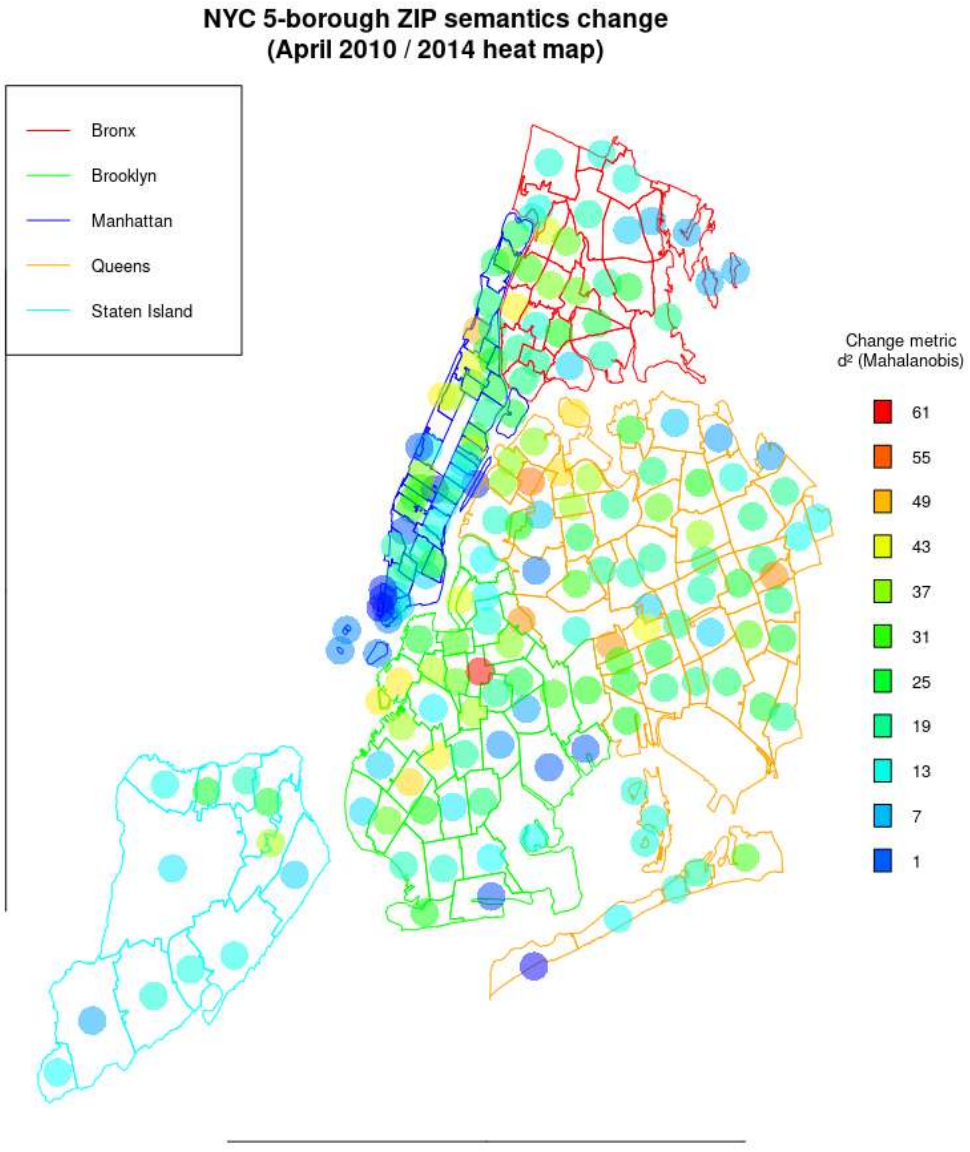


***Figure 28a**:Topographical representation of the intensity of change based on the squared Mahalanobis distance between observed ZIP code areas across the time window: April 2010 ➝ April 2014.*

In Fig. 28a above, the areas of greatest change are central Brooklyn, followed by four orange-colored areas in Queens.

Areas registering the least change may not be quiet areas or areas generally characterized by a reduced number of SRCs or CRCs. Instead they may be areas which consistently exhibit a large proportion of complaints and crime reports.

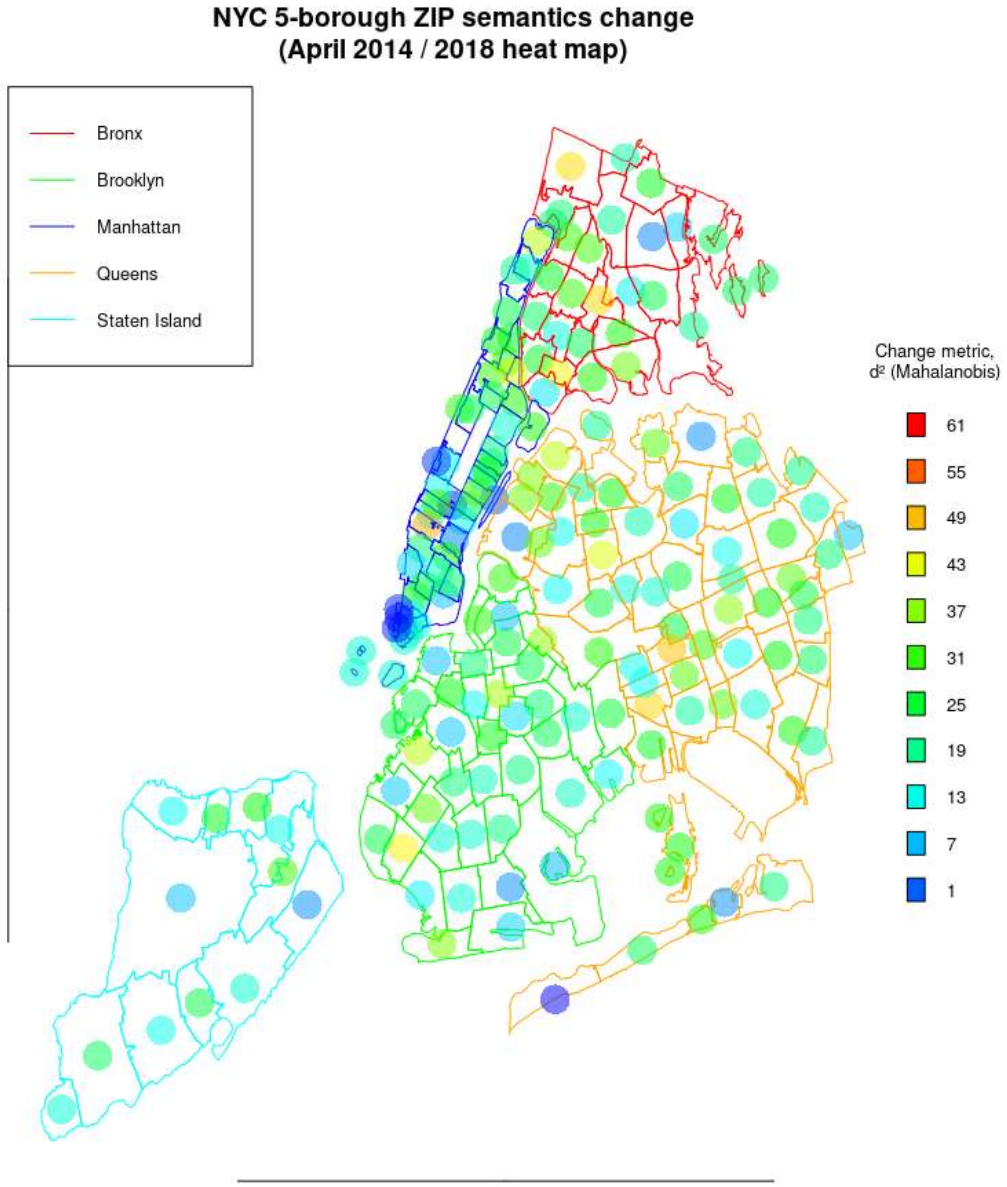


***Figure 28b**: Same as in Fig. 28a, for the time window: April 2014 → April 2018.*

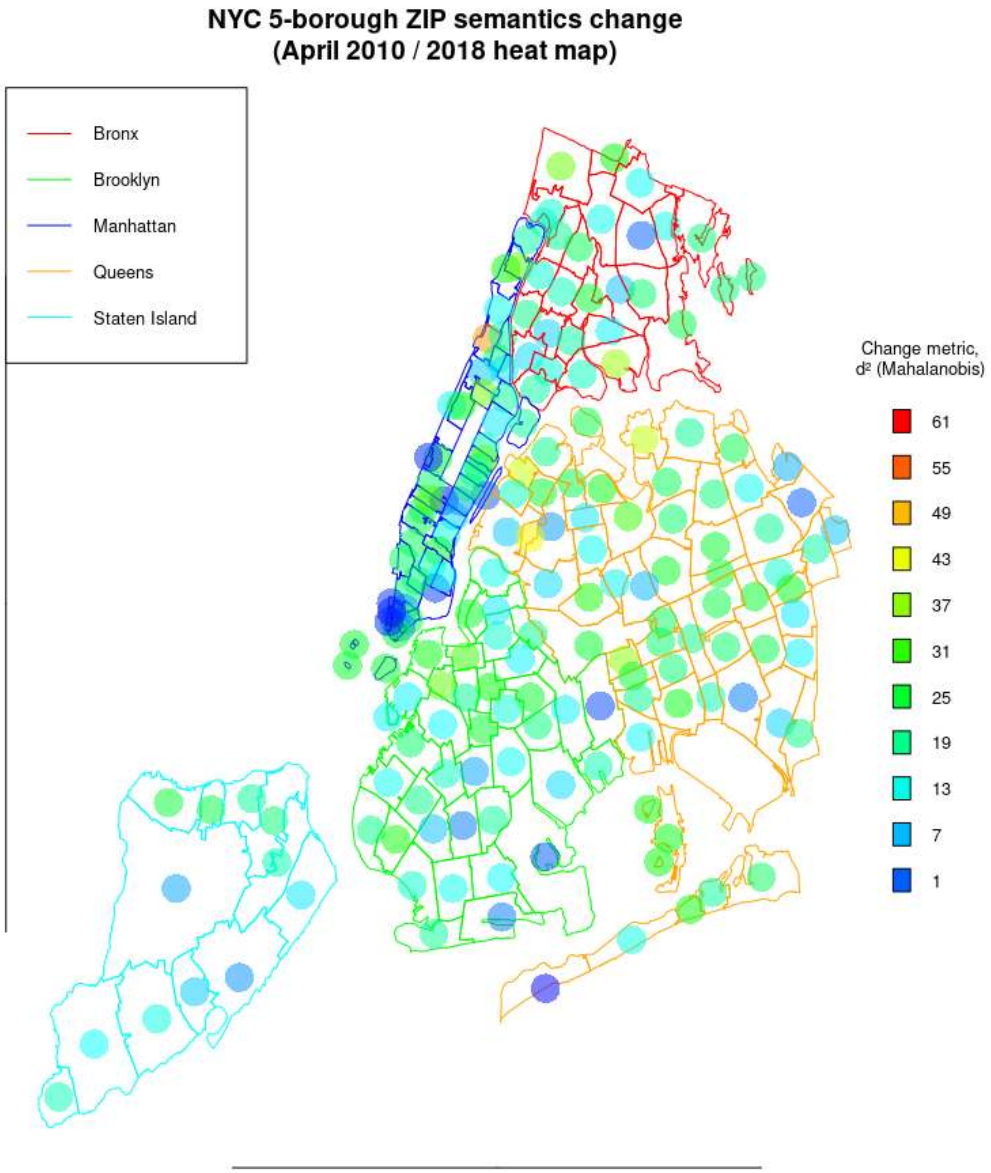


***Figure 28c**: Same as in Fig. 28a, for the time window: April 2010 → April 2018 (not taking into account April 2014 observations).*

Although easy to read, heat maps calculated as above, contain limited information, and are difficult to integrate in a fast update cycle for temporal evolution visualization. They point toward the need for improved visualization tools, capable of unequivocally representing multidimensional data and its temporal evolution in such a way that a non-specialist could readily embrace.

5. Extending this work: a research proposal

From the point of view of the data they generate, large urban areas are complex environments. They are characterized by a large, heterogeneous set of co-varying variables, whose modalities coalesce in space and time as singular events. When put together, those events constitute the unfolding components of a multi-threaded *urban narrative*. We propose to explore possible extensions to the work presented in Sections 1 to 4 of this report. , in a way consistent with the exploitation of heterogeneous spatial data and time series, collected by large metropolitan areas such as New York City, Barcelona or Paris.

This section is made of extracts from a Research Proposal [8] submitted recently for funding to the BBVA Foundation in Spain.

5-1. Introduction

Our proposed extension, "***visualCity***", short for "Visualizing the Mood of Cities" addresses the need for visualizing complex high-dimensional data in a way accessible to most. In the context of smart cities, urban data may typically be made of several hundreds thousands observations a month, across 100 or more dimensions. The project's ultimate objective is to empower end-users (i.e. private citizens, businesses, city governments officials), to promote data transparency and citizen participation in the governing process and to meet stakeholders' objectives in terms of big data utilization, accountability and utility.

As already pointed out, heterogeneous urban data consists of data with vastly different update rates and differing similarity measures. From an update rate viewpoint, we distinguish:

- low-latency, i.e. high update rate data (weak signals), e.g. picked from logged citizen service request calls, complaint calls to the municipality, adequately parsed social media data, crime data, etc., and

- high-latency, i.e. low update rate data (strong signals, e.g. picked from geolocalized census data, income, unemployment or academic achievement data, etc).

From Section 1 of this report, typically, low-latency data has update rate in the range 1 second to 50 hours. By contrast high-latency data has update rates upward of 50 hours and up to 10 years. The 50 hour limit is somewhat arbitrary and meant to fit domain specific data collection mechanisms.

By means of non-linear multidimensional scaling, we propose build a set of analysis tool based on open-source, geo-localized, high and low-latency data. In keeping with its end-user and stakeholder centered approach and objectives, a central requirement is that data ought to be *visualized* in a robust and user-friendly way.

The proposed extension is meant to span a period of 12 months and draws on at least three specialty areas:

(i) urban science,
(ii) data science and statistical learning,
(iii) visualization via computer graphics, as applied to urban data [9].

The projected visualization layout in (iii) should show information on a topographic map, with tools for querying the data by visual brushing and filtering through the use of direct manipulation over the map. Widgets, such as sliders and drop-down lists would entitle users to express simple queries in the familiar environment of a web-based graphical user interface (GUI).

At project's extension's end, the visualization platform, previously developed based on a particular city's urban data, should be tested with distinct urban data gathered in altogether different urban areas.

5-2. Objectives

The remit of Project ***visualCity*** clearly fits against a background of growing political, social and economic interest. That

interest is well exemplified by on-going urban data based projects at the *Urban Dynamics Lab*[ix] at University College London, and at *UBDC*[x] *of* the University of Glasgow, two academic research centers, funded in part by UK's Economic and Social Research Council.

5-2-1. Aim and short-term objectives

Motivated by increasingly large volume of heterogeneous data generated by our urban environments, ***visualCity***'s aim is to provide visual and interpretative help to city officials, citizens and businesses, for them to correctly grasp urban trends, city hot-spots, and pathologies, their expressions, how they correlate, as well as their evolution across the urban landscape and over time. In this context, ***visualCity***'s objectives are:

- **To setup a framework for visualizing, analyzing, and learning, large sets of heterogeneous, co-varying, high dimensional urban data**, and to provide a **dimensionality reduction** which is both domain specific and consistent with end-user expectations. Practical challenges include: (i) integrating heterogeneous data with vastly different update rates into a continuously updated statistical and machine learning method pipeline, and (ii) integrating several measures of proximity: in space, over time and via the semantic description of singular urban events.

- **To propose a functional visualization tool**, which transparently builds on machine learning analytical results, and data high-dimensionality reduction methods, while (i) being user-friendly, (ii) creating a sense of ownership in users by empowering them, (iii) facilitating learning, and when applicable, (iv) assisting users in their decision-making process.

As stated earlier, ***visualCity***'s objectives are tailored to fit a one year development program. Results should include a fully functional programmatic framework as proof of concept, aimed at validating our approach on real data. Data coming from Barcelona Open Data justifies that special care be applied to integrate our framework in *cityOS*[xi], a data platform currently used by the Barcelona municipality. This does not preclude our interest in testing such a visualization platform with extraneous data, e.g. data from cities such as of New York, at project's end.

5-2-1. Long-term objectives

We expect Project ***visualCity*** to be expanded further at an even later stage, i.e. outside the research calendar of this proposed extension. Avenues for expansion include: (i) new algorithms optimized for speed, (ii) a greater degree of automation in the treatment of outliers, (iii) a wider input data spectrum which implies a degree of data interoperability, an improved input interface automation, and (iv) regression analysis to link weak and strong signals so weak-signal-based *what-if* scenarios may be implemented and added to end-users' query toolbox.

5-3. Project situation

5-3-1. Background and state of the art

As urban environments become denser and more complex, in particular in the context of widespread data acquisition in smart-cities, ever increasing amounts of heterogeneous data are generated and collected at different rates. This severely limits the practical applicability and interpretability of statistical analysis without dimensional reduction. Reducing data dimensionality makes unveiling hidden data structure easier and cheaper by improving both methods' runtime performance and results' interpretability at visualization time. In large urban settings millions of weekly observations over hundreds of quantitative variables and categorical variables' modalities representing as many dimensions, make dimensional reduction combined with domain-specific knowledge especially attractive.
Our previous exploratory work (sections 1 to 4) on urban data from New York City (NYC) provides evidence that a conventional statistical learning approach does provide insight in the spatial and social implications of crime and calls generally requesting assistance from municipal services. However it is not always enough to capture all the data's inertia or to visualize results adequately. In typical fashion, our approach was based on linear multi-dimensional scaling, to reveal hidden data structure by feature extraction and factorial analysis in principal components' space. Traditional multivariate

ix https://www.ucl.ac.uk/bartlett/casa/research/current-projects/urban-dynamics-lab
x https://www.ubdc.ac.uk/research
xi https://ajuntament.barcelona.cat/digital/en/digital-transformation/city-data-commons/cityos

analysis (MVA) methods used for high dimensionality reduction include PCA, Singular Vector Decomposition (SVD), Correspondence Analysis (CA), possibly augmented by Varimax, Multiple Correspondence Analysis (MCA), all linear techniques. Those technique are useful in identifying directions of maximum variance and in removing noise, i.e. measured/ collected data fluctuations not explained by covariance analysis. Classical multidimensional scaling (MDS) [10] is another linear method based on a distance matrix in Euclidean space whose treatment with SVD is close to PCA, but whose runtime scales as $O(n^3)$. More scalable refinements based on classical MDS were developed later [11], and all take an initial observations' distance matrix as input and produce vectors in a Euclidean space of arbitrary (lower) dimension, a process generally known as *embedding.*

We previously illustrated how typical results from conventional statistical learning methods, applied to New York City urban data (SRCs to 311 and/or CRCs to 911), include factorial planes projections of observations and feature vectors. Similar projections may be augmented with cluster class information derived from hierarchical clustering (HC) with k-means consolidation. Applying the above conventional methods successively to data defined over successive time windows permits the rudimentary visualization of temporal evolution of urban characteristics for each observed ZIP code.

Except for HC and k-means, all the methods described above are basically matrix factorization technique. They have in common the fact that low dimensional representations keep dissimilar high-dimensional points far apart. On the contrary, in an urban context, we should favor preserving the short to medium range data structure of high dimensional data points. Those points are our multi-dimensional observations. They initially lie on the surface of a hyperplane, i.e. a subspace whose dimension is equal to the number of significant variables explaining the data. In the general case, the hyperplane, technically referred to as a *(sub-)manifold*, is not linearly or quasi-linearly embedded in observation space [12]. Those non-linear (sub-)manifolds are objects which cannot be linearly mapped onto a 2 or 3 dimensional visualization space, subject to the requirement of similarity measure conservation. Urban data analysts and end-users should therefore consider non-linear dimensional reduction (NLDR) schemes, also known as manifold-learning, such as *UMAP* [13], *t-SNE* [14], [15], to preserve data structure on different scales.

After low dimensional embedding with *t-SNE*, the original input features are not explicitly available anymore, although the output of such methods can be used as input for further classification and clustering. Unlike *t-SNE, UMAP* is a parametric manifold-learning method, reported to not only preserve short to medium range similarity (i.e. distance between data points in high dimensional space), but also the long range similarity [13]. Being a parametric method, it permits training the learned model on historical data before submitting test data to it (i.e. new input data) as it becomes available. This results in a reported computational efficiency for the very recent *UMAP* algorithm superior to that of *t-SNE [16]* at the cost of a more complex implementation.

Our long-term objectives and the somewhat natural extension of this proposal is to englobe more diverse urban data as input acquired on vastly different time scales [17]. A wide-ranging body of literature exists which deal with data diversity issues. However a great many reports across the SmartCities research field opt to highlight data generating devices' interoperability [18], rather than data itself, as a key limiting factor in the development of smart interconnected cities. As urban data analytics remains our primary interest, we adopt a *data-centric* approach, aware of the fact that a deluge of data does not guarantee precision, and does not ensure that the choice of proxy variables to represent abstract concepts such as "poverty" or "mobility" is appropriate. Any future extension to this work will pay particular attention to the difficulties of:

- performing cross-cutting analytics where urban data-sets are diverse in nature and origin[xii],
- interpretability, uncertainty quantification and selection bias [19],

when dealing with and pooling urban data.

5-3-2. Applicability and relevance

Increasing average urban populations worldwide

According to the United Nations Population Division[xiii], current average worldwide urban population accounts for more than 55% of all people on Earth. This statistics hides great disparities between countries as well as within countries (e.g. as in the People's Republic of China). Belgium currently stands at 98%, Australia and Brazil at 86%, Argentina and Japan at 92%,

xii Examples of diverse datasets on big data: UBDC at http://ubdc.gla.ac.uk/dataset, BCN Open Data at http://opendata-ajuntament.barcelona.cat/data/en/dataset, NYC Open Data at https://opendata.cityofnewyork.us/data.

xiii World Urbanization Prospects: 2018 Revision, available at https://data.worldbank.org

China at 58% and Egypt at 43% (but oddly falling!). Extrapolating trends observed for the past 15 years, the worldwide average urban population should reach at least 60% in 2030 and close to 90% for industrialized nations alone.

This tells us a great deal about the upcoming challenges posed by increasing urban population densities and the pressure exerted by urban citizens on city governments, so city inhabitants may enjoy minimum standards of quality of life and safety. Those translate as issues and challenges for all: city governments, businesses, and the public at large. It also makes the need for urban data analytics and related visualization tools particularly obvious.

Exploitation of urban data

This work's general contributions goes toward exploiting urban open data, and applying advanced learning and visualization techniques. Project ***visualCity*** provides tools to tackle complex urban planning issues, as well as the related uncertainty suffered by business and the greater public, subtended by hidden multivariate mechanisms and intricated correlations.

Data driven solution

From project ***visualCity***'s inception, we propose a data driven solution, which should vastly improve the visualization of multidimensional urban data by all audiences, anywhere, and in a way that is readily customizable as a function of available data.

Domain expertise

Two parties are included as participating or advisory members in this project, who are expertly aware of urban issues:

- Office of Open Data at the Barcelona City Hall (Ajuntament de Barcelona, Spain).
- NYU Center for Urban Science and Progress (CUSP), a degree-granting technology and research institute in New York City.

Not only will they provide access to Open Data, they will also provide key domain-specific knowledge as to the applicability and pertinence of future results.

Interactivity, usability and testing

The visualization mantra we will adopt explicitly promotes interactivity and usability as a way to empower end-users and encourage the direct use of analytical results, obtained from advanced non-linear multi-dimensional scaling and visualization methods.

A functional platform, available at or very near project's end should allow testing and evaluation by users and stakeholders. That platform is meant to be integrated in cityOS, the production environment of the Office of Open Data of Barcelona.

Scope of queries

- Based on the collected users' feedback, we hope to be able to expand ***visualCity*** to include linear and non-linear regression methods at a later stage. From an in-depth exploratory analysis of selected low-latency data, we expect weak signals to reveal trends and help in the prediction of future evolutions and urban strategies' outcomes. This project's extension points squarely in the direction of *what-if* scenarios in a very large number of use cases pertaining to business and public administration.

- Farther away in the future still, we wish to implement a diversification in input data, a broader set of variables and their suitable treatment. We envision an improved visualization tool integrating text, pictures and rich semantics analysis.

5-4. Research methodology

We go beyond a linear statistical learning approach. We speculate that we can improve on classical low-dimensional representations of results and their simpler non-linear extensions such as hierarchical clustering. What follows is a methodology for further analysis and modelling.

In spite of being a non-convex data visualization heuristics likely to yield local maxima, *t-SNE* has become one of the few *de facto* standards in data visualization based on non-linear manifold learning [15]. We therefore choose to focus this proposal primarily on *t-SNE*, leaving *UMAP* aside in this first speculative phase.

We intend to modify the form data input and use a dimensionless, normalized set of component values for each observation, to include observation time, as well as data points' spatial coordinates. Given the modification of the *t-SNE* method to include new observations coordinates and a proper choice of time normalization, we further propose to implement a Bayesian predictive inference mechanism, where data points' similarity matrix in the original data-set (high dimensional space) is iteratively calculated based on a Bayesian update mechanism, restricted to the new input data. The distinct advantage of such an approach is to be computationally scalable as it does not require the re-computation of all pairwise similarities to carry out an update of the visualization.

▪ We first implement the *t-SNE* method to visualize the more relevant short range structure in our data. "Short range" may be construed simultaneously in three ways: proximity of events in space, in time, and in terms of their categorical modalities and quantitative attributes.

▪ Second we modify *t-SNE* in two ways.

- We consider our geolocated urban data as the set of real-valued random variables $f_{i\in\{1,N\}}=(f_{i1},\ldots,f_{ip})$, for *N* observations. Data points' coordinates are a time-ordered and normalized set of coordinates $f^{(t)}{}_{i\in\{1,N\}}=(f_{i1},\ldots,f_{ip},u_i,v_i,t/\tau_i)$ where:

  - time *t*'s normalization constant, τ_i, is observation-dependent when made to depend on the conformation (curvature and path length during the time step and along the trajectory) of each observables' time trajectory in *p*-dimensional factorial space,

  - *u,v* are simply the observations' geo-location coordinates.

- We reformulate the conditional probability used in calculating the similarity measure in high dimensional space, based on Bayes' theorem. We differentiate between old data, D_{old}, and new data, D_{new}. D_{old} consists of strong signal and past weak signals, before current data update. D_{new} mainly refers to new data brought in by the current update and always consists of weak signals characterized by a high update rate. In order to represent data points' similarity in high dimensional space, we now assume that a true position exists for each observed data point, x_i^o. We assume that newly input data originating from weak signals are distributed as small multivariate Gaussian perturbations, e_i, about the true position, x_i^o. Following we observe that $\{ x^0_i=x^0_j\}$ is equivalent to $\{e_i-e_j = x_i-x_j\}$. If our perturbations, e_i, are independent, the appropriate choice of covariance matrix for the multivariate Gaussian perturbation distribution corresponds to a diagonal matrix [20]. Conveniently the random variable e_i-e_j would also follow a multivariate gaussian distribution. A complete discussion is not warranted at this time but will no doubt require some scrutiny. We can write:

$$P(x_i^o=x_i^o|D_{old}+D_{new})=P(D_{new}|x_i^o=x_i^o,D_{old})\cdot P(x_i^o=x_i^o,D_{old})/P(D_{old}+D_{new})$$

  The prior $P(x_i^o=x_i^o,D_{old})$ is the probability distribution of the similarity between data points before new data points are brought into play (before current data input). It is subject to the update procedure.

  The likelihood $P(D_{new}|x_i^o=x_i^o,D_{old})$ indicates the compatibility of the current (new) data input with the hypothesis. A simplification for the likelihood term, consists in assuming conditional independence of D_{new} and D_{old} given $x_i^o=x_i^o$. In doing so we shall consider the likelihood to be well approximated by $P(D_{new}|x_i^o=x_i^o)$.

  The posterior, or left hand side term in the above, is the probability of the similarity hypothesis based on the complete data (old and new), including current input.

  The denominator on the right hand side is known as the evidence. It can be viewed as a normalization quantity and plays no role in our iterative procedure.

Each new data input triggers a prior's update, where the new prior assumes the value of the last posterior. Predictive Bayesian inference consists in computing a posterior predictive function, as the distribution of new data points marginalized over the prior, i.e. a new likelihood. A partial recalculation of the conditional pairwise probabilities follows, used in evaluating the Kullback-Leibler divergence between similarity measures in the original data set and the embedded representation in low dimensional space. At this point *t-SNE*'s algorithm resumes.

We believe that the above represents an interesting angle of attack for an extension to this preliminary work, aimed at obtaining a scalable, continuous and interactive visualization platform for multidimensional urban data.

Both *t-SNE* and *UMAP* are generic force-directed neighbor graphs techniques. Should our modified *t-SNE* treatment of new data inputs bring us insufficient speed gains for convenient visualization of input data updates, or should it offer mitigated results on heterogeneous urban data, other techniques, combined to *UMAP*, such as *Random-Projection Trees* [21] and *Nearest-Neighbor Descent [22]* are available to rapidly find approximate nearest neighbors in high dimensional space. Meanwhile *Stochastic Gradient Descent* with momentum updates [23] responds to the need for an efficient optimization of the low-dimensional embedding layout. In all cases, either C++ or Python combined with Numba for compiling can be used.

6. Conclusions

The objectives of this exploratory work performed on New York City urbn data were to show:

(a) how much of the the hidden structure of complex urban data can be revealed using conventional multivariate statistical methods,

(b) how trends in the development, disappearance or displacement of urban issues can be extracted from the systematic study of successive time periods,

(c) how well multidimensional data consisting of data request calls enriched with crime reports from NYC inhabitants can be visualized in low dimensional space (mostly 2D).

This was not only to try to understand crime and its statistical correlations with other weak signals in an urban context. Ultimately the larger goal was to propose a research thrust designed to help urban decision-makers and administrators assign budgetary and human resources based on more reliable data models and efficient data visualization.

The composite data set used in this work is formed by data originating from different digital sources. Data pre-processing and generally speaking ETL at data mining stage occupied about 65% of our time and required more than 5000 lines of R code. We produced an automated ETL pipeline capable of processing complex, composite data almost unattended. Only the the detection of outliers could not be satisfactorily automated and required human assessment in its last stage. It is because the final criterion for designating outliers often relies on domain-expertise and not only on quantitative testing. For instance observations (ZIP codes) "10463" (Riverdale in the Bronx) and "11430" (JFK Airport in Queens) are meaningful observations (or "row-profiles") and at the same time outliers at least during the April 2014 period. When included, they exert considerable influence on analytical results overall.

Data mining revealed that data is all too often incomplete, sometimes wrong, or simply statistically unreliable. This was the case for the continuous variable `jlBenef`, short for "jobless benefits" (source: IRS), which proved patchy at best, and was disregarded beyond the ETL stage. Meanwhile the other imported continuous variable `medianInc`, short for "median income" (source: IRS) was kept for later forays. Statisticians are keenly aware that skewing data, introducing bias(es) is a *caveat emptor*. Caution was applied not to introduce data biases prior or during the analysis, for instance by truncating data. At various stages of the ETL phase, statistical independence tests relying on the simple χ^2 statistics were conducted to put our choices in perspective.

In the first half of our data exploration, we showed how we gain limited information from classical approaches such as CA / PCA and even MCA. Data structure subjected to dimensionality reduction was observed. Based on the determination of directions of maximum variance coupled with feature selection and extraction, latent semantics were proposed at different stages of the work. We confirmed a well-known result, namely that low frequency cells may have a dramatic impact, in particular on the projection in factorial space of CA / PCA results. This led us to gradually and carefully rid our data set

from such spurious effects. One dramatic effect of the addition of crime rate was that it shifted the focus away from Manhattan and onto other boroughs, in terms of overall variance explanatory power.

In the second half or this preliminary study, we deployed generic tools of clustering (unsupervised statistical learning). Our approach was based on the exact same data set as before (April 2010, 2014 and 2018), NYC 311 SRCs augmented by NYPD 911 CRCs. Our latent urban data analysis results, at the end of Section 3, point toward 5 cluster classes for April 2014. Meanwhile we report 6 or 7 cluster classes for April 2010 and 3 or 5 for April 2018, respectively in Appendices F and G.

Cluster classes did not coincide with the geographical limits of NYC's five boroughs, but rather with particular traits of the local geography and of the residents' socio-economic makeup inside those boroughs. The shifting picture of cluster classes as a function of observed period was more difficult to grasp. Variables and intensity levels of those variables' modalities, instrumental in the statistical construction of cluster classes, could be elucidated independently for each observation periods. However to infer a plausible mechanism for the variation of observed cluster classes was not possible, relying on the visualization tools we summoned.

All "statistical events" are neither equally acceptable from a social perspective, nor necessarily equal from a statistical viewpoint. In particular to discuss their relative importance or statistical weight becomes important, when their joint analysis aims at designing decision-making support tools. A brief discussion on the merits of (non-)uniform event weighting was included and led us to aggregate SRCs and CRCs data at ZIP code area level, with uniform weighting. That however is strictly equivalent to weighting cluster classes according to their membership or boroughs according to how many observed ZIP code areas they contain. We are aware that this may also introduce a degree of bias, which at this juncture remains unqualified. Again, this issue was knowingly left unattended in this work.

The temporal analysis of trends, presented in Section 4, allowed us to exhibit changes having taken place between the months of April 2010, 2014 and 2018. We could not conclude whether the partial reversal over 2014–2018 of the trend first witnessed over the period 2010–2014 is a data artifact or corresponds to reality on the ground. Further, heat-maps based on the Mahalanobis distance between the observed row profiles of any ZIP code area over two different time periods were only informative in terms of intensity of change. We conclude that greater attention should be paid to result visualization, in terms of tool development for low dimensional embedding.

In order to further strengthen our multivariate analysis, we would welcome new data sources, with:

- population density and distribution by segments,
- number of hospital beds or of family physician per 10,000 inhabitants within a given radius,
- reliable income distribution data,
- academic achievements levels

This would allow us to embark on predictive classification or regression, for instance on continuous income level variable. That in turn would be usefully supplemented by a measure of quality, based not on a traditional confusion table, but on a multi-target (or -output) regression method such as, as an example, support-vector-machine (SVM) based regression.

As they stand, exploratory results presented in this document could also benefit from a comparison with visualization in clusters obtained with the more diversified toolkit of Machine Learning. As our primary objective is ultimately to further our understanding of complex multi-dimensional urban data for dense metropolitan areas such as New York City, or Barcelona, our plan for extending this work is to establish a visualization framework according to a three-pronged approach:

1) Data Extraction, Transformation and Loading (ETL) phase

Automation of the ETL pipeline (data cleaning, outlier detection, feature extraction), and types of data at hand:

- data with low update rates (strong signal).
- data with higher update rates (weak signals).

This preparatory aspect of the project is already largely covered the preliminary work described in Section 1 of this report.

2) Analysis, modeling and results' preliminary representation phase

We modify a manifold-learning technique (be it t-distributed Stochastic Neighbor Embedding or UMAP) to include time and spatial coordinates of observables with the proper normalization. In doing so the embedded (low-dimensional) clustered representation of the original high-dimensional data may usefully accept temporal and spatial distance as an additional component of data points similarity. In the case of the non-parametric method *t-SNE*, a Bayesian update mechanism of the conditional probability used in the computation of the similarity measure in the high dimensional observations' feature space, should provide an approximate but computationally efficient path for rapid visualization updates. The parametric

method, *UMAP*, may further improve the visualization update rate. In either case the goal is to permit the rapid input inclusion of new data (weak signals input) and to mitigate or obviate the need for re-computing past history. As the proposed mechanism will permit to visualize the co-occurrence of urban events in space and time, results in this area pave the way for the almost-continuous visualization of urban events evolving over time.

3) Results' visualization phase

This phase involves the development of a user-friendly web-based interface for end-users to be able to interact graphically and explore the results or our analyses. By making our results accessible via a web interface, the amount of users who will benefit from ***visualCity*** and leverage otherwise somewhat complex analytical results is bound to increase dramatically. Technically oriented users, as well as businesses, city government decision -makers or curious citizens will all be empowered to understand and use the gathered and processed data.

One core difficulty, during the visualization of results obtained from learning methods and high dimensional data analysis, is how to perform rapid visual updates of results as new data becomes available. The fact that urban data aggregates may consist of data with vastly different update rates compounds that difficulty. Another issue, inherent to the learning technique we picked *a priori*, is that *t-SNE* is a non-parametric manifold-learning method. Embedding new data signifies that each augmented data-set should give rise to a completely new low dimensional embedding computation. To address that issue, we recently submitted an exploratory research program dubbed ***visualCity*** for funding. In it we hypothesize the applicability of predictive Bayesian inference to treat the input of new low latency data (weak signals) in order to lower the computational costs of *t-SNE*.

visualCity differentiates between high- and low-latency data (i.e. strong and weak signals respectively). It proposes the implementation of a Bayesian update procedure to calculate the similarity of data points in the initial high dimensional (observation) space in an efficient manner. Our proposed update procedure relies on a multivariate Gaussian distribution of perturbations to represent new low-latency (high update rate) input data against already existing data. When integrating this approach in the *t-SNE* procedure for non-linear MDS, we hypothesize that the perturbations in each dimension are independent of one another to some order and are conditionally independent with respect to past and future perturbations. We therefore surmise that a computationally scalable recalculation of the Kullback-Leibler criterion should result. We further expect solutions with closed analytical formulations, obviating the need to resort to a very large number of Monte Carlo simulations to calculate probability distributions and expectations by averaging them. We expect that the above approach should translate as speed gains when recalculating a new visualization after input of new weak signals (i.e. new low-latency input data).

▪ ▪ ▪

REFERENCES

[1] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

[2] H. F. Kaiser, "The varimax criterion for analytic rotation in factor analysis.," *Psychometrika*, vol. 23, pp. 187–200, 1958.

[3] @amoeba, "PCA - How to compute varimax-rotated principal components in R?," *Cross Validated*. Oct-2017.

[4] "k-means clustering," *Wikipedia*. 20-Oct-2018.

[5] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, "Cluster Analysis: Basic Concepts and Algorithms (Chapt 8)," in *Introduction to Data Mining*, 2nd ed., 1 vols., U. Minnesota, 2019, pp. 487–568.

[6] H. Abdi and D. Valentin, "Multiple Correspondence Analysis," in *Encyclopedia of Measurement and Statistics*, Neil Salkind (Ed.)., vol. 2, 3 vols., Thousand Oaks, CA: Sage, 2007, pp. 651–656.

[7] J. P. Benzécri, "Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire," *Cahiers de l'analyse des données*, vol. 4, no. 3, pp. 377–378, 1979.

[8] C. K. Bhihe and J. García-Vidal, "Research Proposal to the BBVA Foundation: Visualizing the Mood of Cities." UPC - DAC, Dec-2018.

[9] H. Doraiswamy, J. Freire, M. Lage, F. Miranda, and C. Silva, "Spatio-Temporal Urban Data Analysis: A Visual Analytics Perspective," *IEEE Computer Graphics and Applications*, vol. 38, no. 5, pp. 26–35, Oct. 2018.

[10] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952.

[11] J. C. Platt, "FastMap, MetricMap, and Landmark MDS are all Nyström Algorithms," Microsoft Research, MSR-TR-2004-26, Jan. 2005.

[12] V. de Silva and J. B. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," *Adv. Neur. Infor. Process. Syst.*, vol. 15, pp. 721–728, 2003.

[13] L. McInnes and J. Healy, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," *arXiv*, Feb. 2018.

[14] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. of ML Res.*, vol. 9, pp. 2579–2605, 2008.

[15] S. Arora, W. Hu, and P. K. Kothari, "An Analysis of the t-SNE Algorithm for Data Visualization," *arXiv:1803.01768 [cs]*, Mar. 2018.

[16] L. McInnes, *UMAP Uniform Manifold Approximation and Projection for Dimension Reduction*. Austin, TX, 2018.

[17] M. Batty, "Big Data and the City," *Built Environment*, vol. 42, no. 3, pp. 321–337, Oct. 2016.

[18] T. Blewitt, "Interoperability: The key to the emerging smart city," *ReadWrite*, Mar-2017.

[19] D. B. Dunson, "Statistics in the big data era: Failures of the machine," *Stat. and Probab. Letters*, vol. 136, pp. 4–9, May 2018.

[20] J. García-Vidal, Personal communication, December 2018.

[21] S. Dasgupta and Y. Freund, "Random projection trees and low dimensional manifolds," UC San Diego, La Jolla, San Diego, CA, UC San Diego - CS&E Department Technical Report CS2007-0890, 2008.

[22] W. Dong, C. Moses, and K. Li, "Efficient k-nearest neighbor graph construction for generic similarity measures," in *Proc. 20th Int'l Conf. on World Wide Web (WWW '11)*, Hyderabad, India, 2011, p. 577.

[23] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *Proc. 30th Int'l Conf. on Machine Learning*, Atlanta, GA, 2013, p. 14.

APPENDICES

Appendix A: data set's variables' dictionaries

NYC 311 Service Request Calls (SRCs) – Raw Data Dictionary

Column Name	Description
Unique Key	Unique identifier of a Service Request Call (SRC) in the open data set
Created Date	Date SR was created Date in format MM/DD/YY HH:MM:SS AM/PM
Closed Date	Date SR was closed by responding agency. Date in format MM/DD/YY HH:MM:SS AM/PM
Agency	Acronym of responding City Government Agency
Agency Name	Full Agency name of responding City Government Agency
Complaint Type	This is the fist level of a hierarchy identifying the topic of the incident or condition. Complaint Type may have a corresponding Descriptor (below) or may stand alone.
Descriptor	This is associated to the Complaint Type, and provides further detail on the incident or condition. Descriptor values are dependent on the Complaint Type, and are not always required in SR.
Status	Status of SR submitted: Assigned, Canceled, Closed, Pending, +… (Prior column indicates most frequent)
Due Date	Date when responding agency is expected to update the SR. This is based on the Complaint Type and internal SLAs. Date in format MM/DD/YY HH:MM:SS AM/PM
Resolution Action Updated Date	Date when responding agency last updated the SR. Date in format MM/DD/YY HH:MM:SS AM/PM
Resolution Description	Describes the last action taken on the SR by the responding agency. May describe next or future steps.
Location Type	Describes the type of location used in the address information
Incident Zip	Incident location zip code, provided by geo validation.
Incident Address	House number of incident address provided by submitter.
Street Name	Street name of incident address provided by the submitter
Cross Street 1	First Cross street based on the geo validated incident location
Cross Street 2	Second Cross Street based on the geo validated incident location
Intersection Street 1	First intersecting street based on geo validated incident location
Intersection Street 2	Second intersecting street based on geo validated incident location
Address Type	Type of incident location information available (Values: Address; Block face; Intersection; LatLong; Placename)
City	City of the incident location provided by geovalidation.
Landmark	If the incident location is identified as a Landmark the name of the landmark will display here
Facility Type	If available, this field describes the type of city facility associated to the SR
Community Board	Provided by geovalidation.
Borough	Provided by the submitter and confirmed by geovalidation.
X Coordinate (State Plane)	Geo validated, X coordinate of the incident location.

Y Coordinate (State Plane)	Geo validated, Y coordinate of the incident location.
Latitude	Geo based Lat of the incident location
Longitude	Geo based Long of the incident location
Location	Combination of the geo based lat & long of the incident location
Park Facility Name	If the incident location is a Parks Dept facility, the Name of the facility will appear here
Park Borough	The borough of incident if it is a Parks Dept facility
School Name	If the incident location is a Dept of Education school, the name of the school will appear in this field. If the incident is a Parks Dept facility its name will appear here.
School Number	If the incident location is a Dept of Education school, the Number of the school will appear in this field. This field is also used for Parks Dept Facilities.
School Region	If the incident location is a Dept of Education School, the school region number will be appear in this field.
School Code	If the incident location is a Dept of Education School, the school code number will be appear in this field.
School Phone Number	If the facility = Dept for the Aging or Parks Dept, the phone number will appear here. (note - Dept of Education facilities do not display phone number)
School Address	Address of facility of incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School City	City of facilities incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School State	State of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept NY
School Zip	Zip of facility incident location, if the facility is associated with Dept of Education, Dept for the Aging or Parks Dept
School Not Found	Y' in this field indicates the facility was not found (Y; N; BLANK)
School or Citywide Complaint	If the incident is about a Dept of Education facility, this field will indicate if the complaint is about a particualr school or a citywide issue. (Y; N; BLANK)
Vehicle Type	If the incident is a taxi, this field describes the type of TLC vehicle.
Taxi Company Borough	If the incident is identified as a taxi, this field will display the borough of the taxi company.
Taxi Pick Up Location	If the incident is identified as a taxi, this field displays the taxi pick up location
Bridge Highway Name	If the incident is identified as a Bridge/Highway, the name will be displayed here.
Bridge Highway Direction	If the incident is identified as a Bridge/Highway, the direction where the issue took place would be displayed here.
Road Ramp	If the incident location was Bridge/Highway this column differentiates if the issue was on the Road or the Ramp.
Bridge Highway Segment	Additional information on the section of the Bridge/Highway were the incident took place.
Garage Lot Name	Related to DOT Parking Meter SR, this field shows what garage lot the meter is located in
Ferry Direction	Used when the incident location is within a Ferry, this field indicates the direction of ferry
Ferry Terminal Name	Used when the incident location is Ferry, this field indicates the ferry terminal where the incident took place.

NYPD 911 Crime Report Calls (CRCs) – Raw Data Dictionary

CMPLNT_NUM	Unique persistent ID for each complaint or Crime Report Call (CRC).
CMPLNT_FR_DT	Exact date of occurrence for the reported event (or starting date of occurrence if CMPLNT_TO_DT exists)
CMPLNT_FR_TM	Exact time of occurrence for the reported event (or starting time of occurrence if CMPLNT_TO_TM exists)
CMPLNT_TO_DT	Ending date of occurrence for the reported event if exact time of occurrence is unknown
CMPLNT_TO_TM	Ending time of occurrence for the reported event if exact time of occurrence is unknown
RPT_DT	Date event was reported to police
KY_CD	Three digit offense classification code
OFNS_DESC	Description of offense corresponding with key code (KY_CD)
PD_CD	Three digit internal classification code (more granular than Key Code)
PD_DESC	Description of internal classification corresponding with PD code; more granular than Offense Description (OFNS_DESC).
CRM_ATPT_CPTD_CD	Crime completion indicator (completed, attempted but failed, interrupted prematurely)
LAW_CAT_CD	Level of offense (felony, misdemeanor, violation)
JURIS_DESC	Jurisdiction responsible for incident. Either internal (Police, Transit, Housing) or external (Correction, Port Authority, etc."
BORO_NM	The name of the borough in which the incident occurred
ADDR_PCT_CD	The precinct in which the incident occurred
LOC_OF_OCCUR_DESC	"Specific location of occurrence in or around the premises (inside, opposite of, in front of, at the rear of)
PREM_TYP_DESC	Specific description of premises (grocery store, residence, street, etc.)
PARKS_NM	Name of NYC park, playground or greenspace of occurrence if applicable (state parks are not included)
HADEVELOPT	Name of NYCHA housing development of occurrence if applicable
X_COORD_CD	X-coordinate for New York State Plane Coordinate System, Long Island Zone (NAD 83) in units of feet (FIPS 3104)
Y_COORD_CD	"Y-coordinate for New York State Plane Coordinate System, Long Island Zone (NAD 83) in units of feet (FIPS 3104)
Latitude	"Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326) "
Longitude	"Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)

IRS Statistics of Income per ZIP code– Raw Data Dictionary

IRS Documentation Guide (year 2014)

Contents

A. Overview

The Statistics of Income (SOI) division bases its ZIP code data on administrative records of individual income tax returns (Forms 1040) from the Internal Revenue Service (IRS) Individual Master File (IMF) system. Included in these data are returns filed during the 12-month period, January 1, 2015 to December 31, 2015. While the bulk of returns filed during the 12-month period are primarily for Tax Year 2014, the IRS received a limited number of returns for tax years before 2014 and these have been included within the ZIP code data.

B. Nature of Changes

The following changes have been made to the Tax Year 2014 ZIP Code data:

• Two new variables have been added for volunteer prepared returns: volunteered income tax assistance (VITA) and tax counseling for the elderly (TCE) prepared returns.

• Five new variables, related to the Affordable Care Act (ACA), have been added to the data: Excess advance premium tax credit repayment, Total premium tax credit, Advance premium tax credit, Health care individual responsibility payment, and Net premium tax credit. Please refer to section F for a complete list of variables and their corresponding names.

C. Population Definitions and Tax Return Addresses

• ZIP Code data are based on population data that was filed and processed by the IRS during the 2015 calendar year.

• State totals may not be comparable to State totals published elsewhere by SOI because of specific disclosure protection features in the ZIP code data.

• Data do not represent the full U.S. population because many individuals are not required to file an individual income tax return.

• The address shown on the tax return may differ from the taxpayer's actual residence.

• State codes were based on the ZIP code shown on the return.

• Excluded were tax returns filed without a ZIP code and returns filed with a ZIP code that did not match the State code shown on the return.

• Excluded were tax returns filed using Army Post Office (APO) and Fleet Post Office addresses, foreign addresses, and addresses in Puerto Rico, Guam, Virgin Islands, American Samoa, Marshall Islands, Northern Marianas, and Palau.

D. Disclosure Protection Procedures

SOI did not attempt to correct any ZIP codes on the returns; however, it did take the following precautions to avoid disclosing information about specific taxpayers:

• ZIP codes with less than 100 returns and those identified as a single building or nonresidential ZIP code were categorized as "other" (99999).

• Income and tax items with less than 20 returns for a particular AGI class were combined with another AGI class within the same ZIP Code. Collapsed AGI classes are identified with a double asterisk (**).

• All number of returns variables have been rounded to the nearest 10.

• Excluded from the data are items with less than 20 returns within a ZIP code.

• Excluded from the data are tax returns with a negative adjusted gross income.

• Excluded are tax returns representing a specified percentage of the total of any particular cell. For example, if one return represented 75 percent of the value of a given cell, the return was suppressed from the tabulation. The actual threshold percentage used cannot be released.

E. File Characteristics

The ZIP code data are available in three formats:

(1) Individual state excel files—14zp##xx.xls (## = 01-51; xx = AL-WY)

(2) A comma separated file (.csv) with AGI classes —14zpallagi.csv

(3) A comma separated filewithout AGI classes(The AGI_STUB variable has been set to zero for this file)—14zpallnoagi.csv

For all the files, the money amounts are reported in thousands of dollars.

F. Selected Income and Tax Items

STATEFIPS	The State Federal Information Processing System (FIPS) code
STATE	The State associated with the ZIP code
ZIPCODE	5-digit Zip code
AGI_STUB	Size of Adjusted Gross Income (AGI) 1 = $1 under $25,000 2 = $25,000 under $50,000 3 = $50,000 under $75,000 4 = $75,000 under $100,000 5 = $100,000 under $200,000 6 = $200,000 or more
N1	Number of returns
...	**...**

G. Endnotes:

For complete individual income tax tabulations at the State level, see the historic table posted to Tax Stats at http://www.irs.gov/uac/SOI-Tax-Stats---Historic-Table-2.

Does not include returns with adjusted gross deficit.

The "Number of volunteer prepared returns" shows counts of returns prepared by IRS-certified volunteers to taxpayers with limited income, persons with disabilities, limited English speaking taxpayers, current and former members of the military, and taxpayers who are 60 years of age and older.

“Qualified dividends” are ordinary dividends received in tax years beginning after 2002 that meet certain conditions and receive preferential tax rates. The maximum qualified dividends tax rate is 15%.

Includes the Alaskan permanent fund, reported by residents of Alaska on Forms 1040A and 1040EZ's.

This fund only applies to statistics in the totals, and the state of Alaska.

Earned income credit includes both the refundable and non-refundable portions. The non-refundable portion could reduce income tax and certain related taxes to zero. The earned income credit amounts in excess of total tax liability, or amounts when there was no tax liability at all, were refundable. See footnote 6 below for explanation of the refundable portion of the earned income credit.

The refundable portion of the earned income credit equals total income tax minus the earned income credit. If the result is negative, this amount is considered the refundable portion. No other refundable credits were taken into account for this calculation.

Income tax reflects the amount reported on Form 1040 line 56. It also includes data from Form 1040A and 1040EZ filers.

“Total tax liability” differs from “Income tax”, in that “Total tax liability” includes the taxes from recapture of certain prior-year credits, tax applicable to individual retirement arrangements (IRA's), social security taxes on self-employment income and on certain tip income, advanced earned income payments, household employment taxes, and certain other taxes listed in the Form 1040 instructions.

[10] Reflects payments to or with-holdings made to "Total tax liability". This is the amount the tax filer owes when the income tax return is filed.

[11] The amount of over-payments the tax filer requested to have refunded.

Appendix B: NYPD's crime categorization

Crime modalities are: felony, misdemeanor, and violation.

FELONY
It is the most serious of offenses and gives rise to a more thorough classification. Felonies are lettered, with Class A being the most serious and Class E being the least serious. They are also divided into a smaller sub category; violent and non violent. In the state of NY, a non-violent, Class D felony would call for 1 to 4 years of probation. However, a violent Class D felony would automatically require a prison sentence of at least 2 years. What characterizes each felony as violent or non-violent is usually the presence of a weapon (possession of a firearm) or bodily harm to another person (aggravated assault/battery). A Class A Felony (e.g a 1st degree murder) is punishable by life in prison, with or without parole, depending on the circumstances.
MISDEMEANOR
Thi second type of criminal offenses is less severe than a felony but more serious than a violation. Misdemeanors can carry up to a year in jail. In addition to jail time, a person convicted of a misdemeanor can also be subject to fines, probation, community service or restitution (victim compensation). A classic case of a misdemeanor would be simple assault, possession of a small amount of marijuana, or driving under the influence.
VIOLATION
Also known as "infractions", it is a minor offense. A speeding ticket, public intoxication, or jaywalking are some of the many petty offenses that could fall under the umbrella of violations. Violations are punishable by fines primarily, and do not result in jail or prison time.

In the subsequent listings, a number following a label within each category indicates the degree of the charge within that category, i.e. sub-categorization for judicial purposes.

Felonies

RAPE 1 *(means "1st degree rape", i.e. generally speaking rape under the threat of a deadly weapon, etc.)*
LARCENY,GRAND BY OPEN/COMPROMISE CELL PHONE ACCT
LARCENY,GRAND BY OPEN CREDIT CARD (NEW ACCT)
RAPE 3
FRAUD,UNCLASSIFIED-FELONY
LARCENY,GRAND BY DISHONEST EMP
BURGLARY,RESIDENCE,NIGHT
SEX CRIMES
RAPE 2
LARCENY,GRAND BY BANK ACCT COMPROMISE-REPRODUCED CHECK
SODOMY 1
LARCENY,GRAND BY THEFT OF CREDIT CARD
LARCENY,GRAND BY FALSE PROMISE-NOT IN PERSON CONTACT
LARCENY,GRAND FROM RESIDENCE, UNATTENDED
SEXUAL ABUSE
LARCENY,GRAND FROM BUILDING (NON-RESIDENCE) UNATTENDED
COERCION 1
PUBLIC ADMINISTRATION,UNCLASSI
COMPUTER TAMPER/TRESSPASS
LARCENY,GRAND FROM OPEN AREAS, UNATTENDED
LARCENY,GRAND BY IDENTITY THEFT-UNCLASSIFIED
BURGLARY,RESIDENCE,UNKNOWN TIM
BURGLARY,RESIDENCE,DAY
LARCENY,GRAND BY FALSE PROMISE-IN PERSON CONTACT
TAMPERING 1,CRIMINAL
RAPE 1, ATTEMPT

LARCENY,GRAND BY CREDIT CARD ACCT COMPROMISE-EXISTING ACCT
LARCENY,GRAND BY BANK ACCT COMPROMISE-TELLER
FORGERY,ETC.,UNCLASSIFIED-FELO
NY STATE LAWS,UNCLASSIFIED FEL
CRIMINAL CONTEMPT 1
LARCENY,GRAND BY BANK ACCT COMPROMISE-ATM TRANSACTION
LARCENY,GRAND BY ACQUIRING LOST CREDIT CARD
MISCHIEF,CRIMINAL, UNCL 2ND
ARSON 2,3,4
RECKLESS ENDANGERMENT 1
MISCHIEF, CRIMINAL 3 & 2, OF M
LARCENY,GRAND OF VEHICULAR/MOTORCYCLE ACCESSORIES
LARCENY,GRAND FROM STORE-SHOPL
LARCENY,GRAND BY BANK ACCT COMPROMISE-UNCLASSIFIED
LARCENY,GRAND BY ACQUIRING LOS
LARCENY,GRAND FROM VEHICLE/MOTORCYCLE
LARCENY,GRAND OF AUTO
BURGLARY,COMMERCIAL,NIGHT
LARCENY,GRAND FROM RETAIL STORE, UNATTENDED
BURGLARY,COMMERCIAL,UNKNOWN TI
LARCENY,GRAND FROM PERSON,PICK
LARCENY,GRAND OF MOTORCYCLE
LARCENY,GRAND BY EXTORTION
WEAPONS POSSESSION 3
FORGERY,DRIVERS LICENSE
LARCENY,GRAND FROM PERSON,PERSONAL ELECTRONIC DEVICE(SNATCH)
ROBBERY,OPEN AREA UNCLASSIFIED
LARCENY,GRAND FROM NIGHT CLUB, UNATTENDED
CONTROLLED SUBSTANCE,INTENT TO
ASSAULT 2,1,UNCLASSIFIED
CONTROLLED SUBSTANCE,POSSESS.
ROBBERY,DWELLING
IMPRISONMENT 1,UNLAWFUL
STRANGULATION 1ST
LARCENY,GRAND FROM EATERY, UNATTENDED
STOLEN PROPERTY 2,1,POSSESSION
LARCENY, GRAND OF AUTO - ATTEM
BURGLARY,TRUCK NIGHT
ROBBERY,PERSONAL ELECTRONIC DEVICE
BURGLARY,UNCLASSIFIED,NIGHT
LARCENY,GRAND OF BICYCLE
ARSON, MOTOR VEHICLE 1 2 3 & 4
WEAPONS POSSESSION 1 & 2
CONTROLLED SUBSTANCE, SALE 5
FORGERY,M.V. REGISTRATION
ASSAULT 2,1,PEACE OFFICER
ROBBERY,COMMERCIAL UNCLASSIFIED
FORGERY-ILLEGAL POSSESSION,VEH
ROBBERY,RESIDENTIAL COMMON AREA
LARCENY,GRAND FROM PERSON, BAG OPEN/DIP
CONTROLLED SUBSTANCE,SALE 1
BRIBERY,PUBLIC ADMINISTRATION
IMPERSONATION 1, POLICE OFFICER
MARIJUANA, SALE 1, 2 & 3

ROBBERY,PUBLIC PLACE INSIDE
MENACING 1ST DEGREE (VICT NOT
CRIMINAL MIS 2 & 3
ROBBERY, PAYROLL
ROBBERY,HOME INVASION
CONTROLLED SUBSTANCE,SALE 3
LARCENY,GRAND FROM PERSON,PURS
THEFT,RELATED OFFENSES,UNCLASS
LARCENY,GRAND FROM PERSON,UNCL
ROBBERY,CAR JACKING
AGGRAVATED HARASSMENT 1
BURGLARY,COMMERCIAL,DAY
LARCENY,GRAND BY BANK ACCT COMPROMISE-UNAUTHORIZED PURCHASE
ROBBERY,POCKETBOOK/CARRIED BAG
CONTROLLED SUBSTANCE, POSSESSI
UNAUTHORIZED USE VEHICLE 2
CONTROLLED SUBSTANCE, INTENT T
BURGLARY,TRUCK DAY
MARIJUANA, POSSESSION 1, 2 & 3
ROBBERY,OF TRUCK DRIVER
CRIMINAL DISPOSAL FIREARM 1 &
CONTROLLED SUBSTANCE,SALE 2
LARCENY,GRAND BY OPEN BANK ACCT
BURGLARY,UNCLASSIFIED,UNKNOWN
FORGERY,PRESCRIPTION
SODOMY 2
GAMBLING 1,PROMOTING,BOOKMAKIN
AGGRAVATED CRIMINAL CONTEMPT
ROBBERY, CHAIN STORE
FALSE REPORT 1,FIRE
ROBBERY,PHARMACY
ROBBERY,LICENSED MEDALLION CAB
STOLEN PROPERTY-MOTOR VEH 2ND,
LARCENY,GRAND OF TRUCK
ROBBERY,LIQUOR STORE
LARCENY,GRAND FROM PERSON,LUSH WORKER(SLEEPING/UNCON VICTIM)
BRIBERY, POLICE OFFICER
ARSON 1
TRESPASS 1,CRIMINAL
ROBBERY,UNLICENSED FOR HIRE VEHICLE
CONTROLLED SUBSTANCE, SALE 4
ROBBERY,BICYCLE
OBSCENE MATERIAL - UNDER 17 YE
ROBBERY,BANK
ROBBERY,NECKCHAIN/JEWELRY
LARCENY,GRAND PERSON,NECK CHAI
ROBBERY,BODEGA/CONVENIENCE STORE
DRUG PARAPHERNALIA, POSSESSION
CUSTODIAL INTERFERENCE 1
ESCAPE 2,1
PROMOTING A SEXUAL PERFORMANCE
BURGLARY,UNCLASSIFIED,DAY
ROBBERY,GAS STATION
MENACING 1ST DEGREE (VICT PEAC

USE OF A CHILD IN A SEXUAL PERFORMANCE
CONSPIRACY 2, 1
SEX TRAFFICKING
INCOMPETENT PERSON,KNOWINGLY ENDANGERING
TAX LAW
MANUFACTURE UNAUTHORIZED RECOR
MISCHIEF, CRIMINAL 3&2, BY FIR
ROBBERY,ON BUS/ OR BUS DRIVER
ROBBERY,ATM LOCATION
LARCENY,GRAND FROM TRUCK, UNATTENDED
OBSCENITY 1
CHILD ABANDONMENT
INTOXICATED DRIVING,ALCOHOL
HOMICIDE, NEGLIGENT, VEHICLE,
MAKING TERRORISTIC THREAT
BURGLARY,UNKNOWN TIME
KIDNAPPING 2
BAIL JUMPING 1 & 2
FACILITATION 3,2,1, CRIMINAL
SOLICITATION 3,2,1, CRIMINAL
END WELFARE VULNERABLE ELDERLY PERSON
AGGRAVATED SEXUAL ASBUSE
LARCENY,GRAND FROM PIER, UNATTENDED
ROBBERY,BAR/RESTAURANT
SODOMY 3
SUPP. ACT TERR 2ND
LARCENY, GRAND OF MOPED
LARCENY,GRAND FROM BOAT, UNATTENDED
SALE SCHOOL GROUNDS 4
KIDNAPPING 1
ROBBERY,CHECK CASHING BUSINESS

Misdemeanors

ASSAULT 3
LARCENY,PETIT FROM BUILDING,UN
FRAUD,UNCLASSIFIED-MISDEMEANOR
AGGRAVATED HARASSMENT 2
SEXUAL ABUSE 3,2
CRIMINAL MISCHIEF 4TH, GRAFFIT
SEXUAL MISCONDUCT,INTERCOURSE
CRIMINAL MISCHIEF,UNCLASSIFIED 4
MISCHIEF, CRIMINAL 4, BY FIRE
MISCHIEF, CRIMINAL 4, OF MOTOR
LARCENY,PETIT OF LICENSE PLATE
CHILD, ENDANGERING WELFARE
UNAUTHORIZED USE VEHICLE 3
VIOLATION OF ORDER OF PROTECTI
PUBLIC ADMINISTATION,UNCLASS M
LARCENY,PETIT BY CREDIT CARD U
CUSTODIAL INTERFERENCE 2
LARCENY,PETIT FROM OPEN AREAS,
NY STATE LAWS,UNCLASSIFIED MIS
LARCENY,PETIT FROM STORE-SHOPL

FORGERY,ETC.-MISD.
LARCENY,PETIT FROM AUTO
STOLEN PROPERTY 3,POSSESSION
LARCENY,PETIT BY FALSE PROMISE
CONTEMPT,CRIMINAL
LARCENY,PETIT BY CHECK USE
BRIBERY,COMMERCIAL
MENACING,UNCLASSIFIED
OBSTR BREATH/CIRCUL
ADM.CODE,UNCLASSIFIED MISDEMEA
LARCENY,PETIT OF VEHICLE ACCES
LEWDNESS,PUBLIC
CONTROLLED SUBSTANCE, POSSESSI
MARIJUANA, POSSESSION 4 & 5
WEAPONS, POSSESSION, ETC
INTOXICATED DRIVING,ALCOHOL
TRESPASS 2, CRIMINAL
THEFT,RELATED OFFENSES,UNCLASS
ACCOSTING,FRAUDULENT
MARIJUANA, SALE 4 & 5
LARCENY,PETIT OF MOTORCYCLE
LARCENY,PETIT OF BICYCLE
RECKLESS ENDANGERMENT 2
LEAVING SCENE-ACCIDENT-PERSONA
IMPERSONATION 2, PUBLIC SERVAN
RESISTING ARREST
TRAFFIC,UNCLASSIFIED MISDEMEAN
LARCENY,PETIT BY ACQUIRING LOS
TRESPASS 3, CRIMINAL
LARCENY,PETIT FROM TRUCK
IMPRISONMENT 2,UNLAWFUL
BURGLARS TOOLS,UNCLASSIFIED
THEFT OF SERVICES, UNCLASSIFIE
LARCENY,PETIT FROM BOAT
LARCENY,PETIT BY DISHONEST EMP
RECKLESS ENDANGERMENT OF PROPE
TAX LAW
UNAUTH. SALE OF TRANS. SERVICE
PETIT LARCENY-CHECK FROM MAILB
IMPAIRED DRIVING,DRUG
ASSEMBLY,UNLAWFUL
BAIL JUMPING 3
FALSE REPORT UNCLASSIFIED
RECORDS,FALSIFY-TAMPER
SEXUAL MISCONDUCT,DEVIATE
PROSTITUTION, PATRONIZING 4, 3
SALE OF UNAUTHORIZED RECORDING
DRUG PARAPHERNALIA, POSSESSE
CHILD,ALCOHOL SALE TO
GAMBLING 2,PROMOTING,UNCLASSIF
CHECK,BAD
FALSE REPORT BOMB
LARCENY, PETIT OF AUTO - ATTEM
RECKLESS DRIVING

AGRICULTURE & MARKETS LAW,UNCL
TAMPERING 3,2, CRIMINAL
PROSTITUTION 4,PROMOTING&SECUR
GENERAL BUSINESS LAW,TICKET SP
LARCENY,PETIT OF BOAT
POSSESSION HYPODERMIC INSTRUME
ALCOHOLIC BEVERAGE CONTROL LAW
GAMBLING, DEVICE, POSSESSION
STOLEN PROP-MOTOR VEHICLE 3RD,
CHILD,OFFENSES AGAINST,UNCLASS
LARCENY,PETIT OF AUTO
PUBLIC SAFETY,UNCLASSIFIED MIS
LARCENY, PETIT OF MOPED
DOG STEALING
DIS. CON.,AGGRAVATED
RIOT 2/INCITING
MENACING,PEACE OFFICER
JOSTLING
PERJURY 3,ETC.
ESCAPE 3
PUBLIC HEALTH LAW,UNCLASSIFIED
COMPUTER UNAUTH. USE/TAMPER
FALSE ALARM FIRE
NUISANCE,CRIMINAL,UNCLASSIFIED
WOUNDS,REPORTING OF
LARCENY, PETIT FROM COIN MACHINE

Violations

HARASSMENT,SUBD 3,4,5
HARASSMENT,SUBD 1,CIVILIAN
MARIJUANA, POSSESSION
ALCOHOLIC BEVERAGES,PUBLIC CON
THEFT OF SERVICES- CABLE TV SE
POSSES OR CARRY A KNIFE
ADM.CODE,UNCLASSIFIED VIOLATIO
PEDDLING,UNLAWFUL
TRESPASS 4,CRIMINAL SUB 2
DISORDERLY CONDUCT
IMITATION PISTOL/AIR RIFLE
PARKR&R,UNCLASSIFIED VIOLATION
NY STATE LAWS,UNCLASSIFIED VIO
APPEARANCE TICKET FAIL TO RESP
IMITATION PISTOL/AIR RIFLE
TRAFFIC,UNCLASSIFIED INFRACTION
LOITERING,GAMBLING,OTHER
ENVIRONMENTAL CONTROL BOARD
INAPPROPIATE SHELTER DOG LEFT
EXPOSURE OF A PERSON
UNDER THE INFLUENCE OF DRUGS

Appendix C: Index of ZIP codes and New York city boroughs

ZIP	Borough
10001	Manhattan
10002	Manhattan
10003	Manhattan
10004	Manhattan
10005	Manhattan
10006	Manhattan
10007	Manhattan
10009	Manhattan
10010	Manhattan
10011	Manhattan
10012	Manhattan
10013	Manhattan
10014	Manhattan
10016	Manhattan
10017	Manhattan
10018	Manhattan
10019	Manhattan
10020	Manhattan
10021	Manhattan
10022	Manhattan
10023	Manhattan
10024	Manhattan
10025	Manhattan
10026	Manhattan
10027	Manhattan
10028	Manhattan
10029	Manhattan
10030	Manhattan
10031	Manhattan
10032	Manhattan
10033	Manhattan
10034	Manhattan
10035	Manhattan
10036	Manhattan
10037	Manhattan
10038	Manhattan
10039	Manhattan
10040	Manhattan
10041	Manhattan
10044	Manhattan
10045	Manhattan
10048	Manhattan
10065	Manhattan
10069	Manhattan
10075	Manhattan
10103	Manhattan
10107	Manhattan
10111	Manhattan
10112	Manhattan
10118	Manhattan

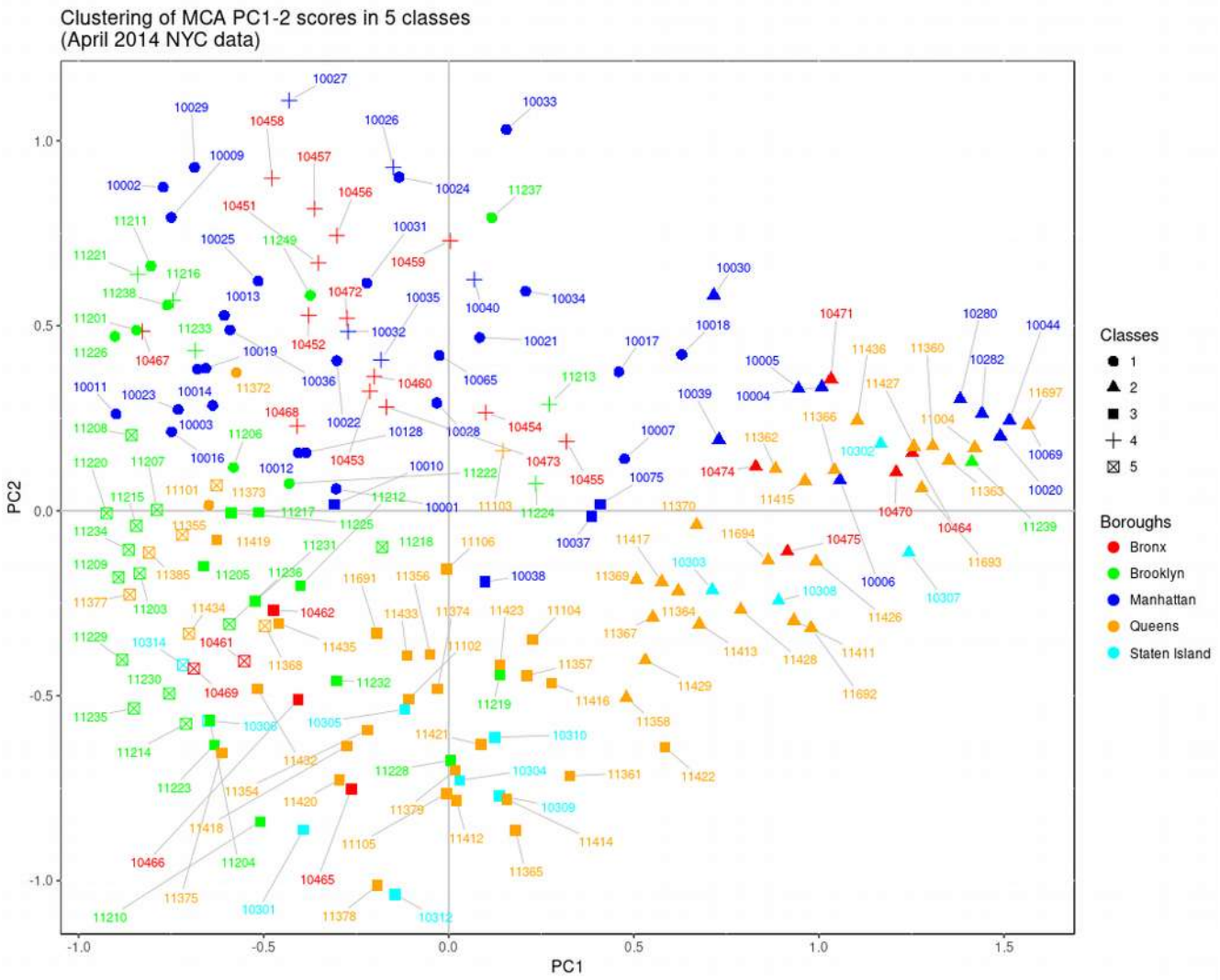
ZIP	Borough
10119	Manhattan
10129	Manhattan
10162	Manhattan
10163	Manhattan
10167	Manhattan
10170	Manhattan
10172	Manhattan
10178	Manhattan
10203	Manhattan
10259	Manhattan
10278	Manhattan
10280	Manhattan
10281	Manhattan
10282	Manhattan
10301	Staten Isl.
10302	Staten Isl.
10303	Staten Isl.
10304	Staten Isl.
10305	Staten Isl.
10306	Staten Isl.
10307	Staten Isl.
10308	Staten Isl.
10309	Staten Isl.
10310	Staten Isl.
10312	Staten Isl.
10314	Staten Isl.
10451	Bronx
10452	Bronx
10453	Bronx
10454	Bronx
10455	Bronx
10456	Bronx
10457	Bronx
10458	Bronx
10459	Bronx
10460	Bronx
10461	Bronx
10462	Bronx
10463	Bronx
10464	Bronx
10465	Bronx
10466	Bronx
10467	Bronx
10468	Bronx
10469	Bronx
10470	Bronx
10471	Bronx
10472	Bronx
10473	Bronx
10474	Bronx

ZIP	Borough
10475	Bronx
11004	Queens
11101	Queens
11102	Queens
11103	Queens
11104	Queens
11105	Queens
11106	Queens
11109	Queens
11201	Brooklyn
11202	Brooklyn
11203	Brooklyn
11204	Brooklyn
11205	Brooklyn
11206	Brooklyn
11207	Brooklyn
11208	Brooklyn
11209	Brooklyn
11210	Brooklyn
11211	Brooklyn
11212	Brooklyn
11213	Brooklyn
11214	Brooklyn
11215	Brooklyn
11216	Brooklyn
11217	Brooklyn
11218	Brooklyn
11219	Brooklyn
11220	Brooklyn
11221	Brooklyn
11222	Brooklyn
11223	Brooklyn
11224	Brooklyn
11225	Brooklyn
11226	Brooklyn
11228	Brooklyn
11229	Brooklyn
11230	Brooklyn
11231	Brooklyn
11232	Brooklyn
11233	Brooklyn
11234	Brooklyn
11235	Brooklyn
11236	Brooklyn
11237	Brooklyn
11238	Brooklyn
11239	Brooklyn
11249	Brooklyn
11251	Brooklyn
11354	Queens

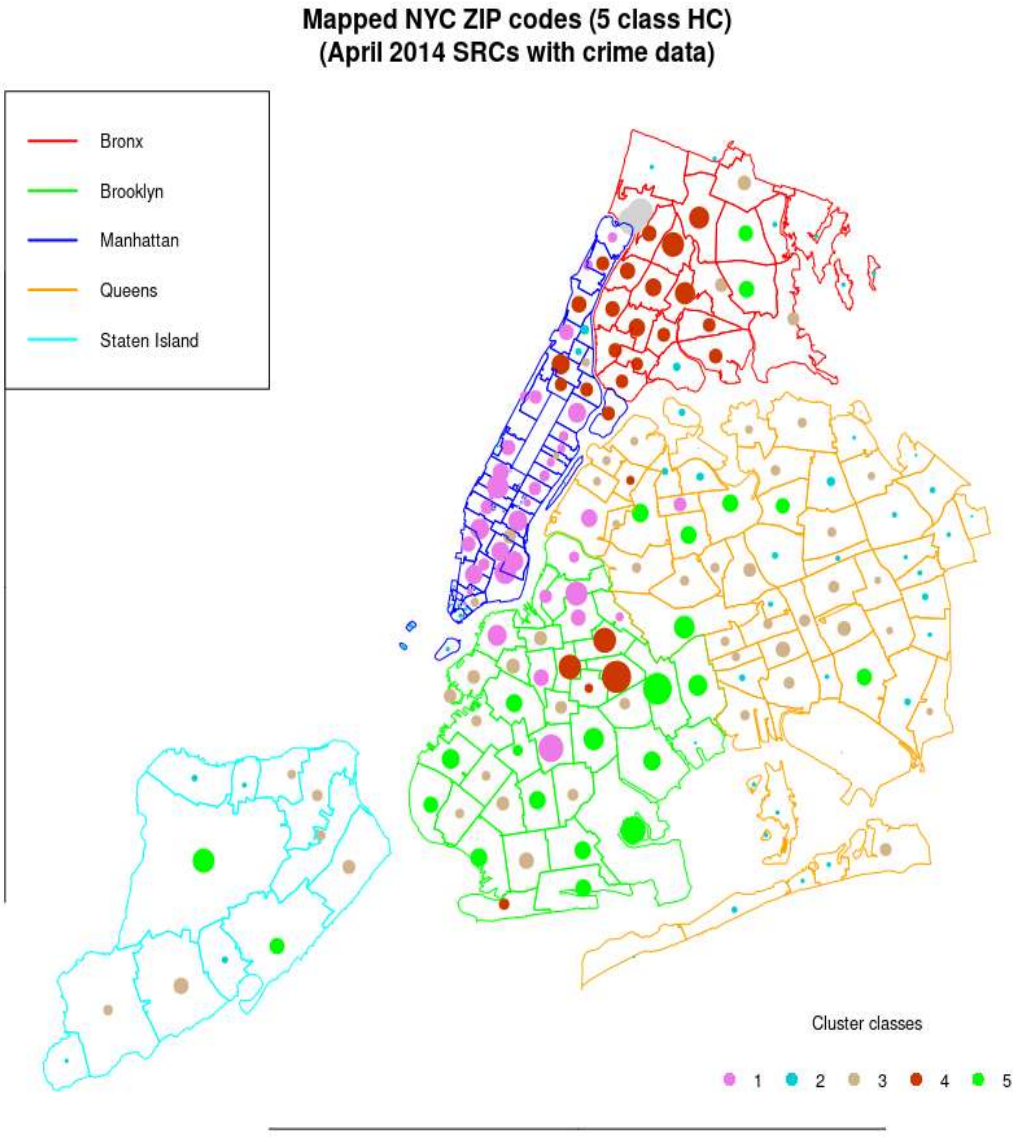
ZIP	Borough
11355	Queens
11356	Queens
11357	Queens
11358	Queens
11359	Queens
11360	Queens
11361	Queens
11362	Queens
11363	Queens
11364	Queens
11365	Queens
11366	Queens
11367	Queens
11368	Queens
11369	Queens
11370	Queens
11371	Queens
11372	Queens
11373	Queens
11374	Queens
11375	Queens
11377	Queens
11378	Queens
11379	Queens
11385	Queens
11411	Queens
11412	Queens
11413	Queens
11414	Queens
11415	Queens
11416	Queens
11417	Queens
11418	Queens
11419	Queens
11420	Queens
11421	Queens
11422	Queens
11423	Queens
11426	Queens
11427	Queens
11428	Queens
11429	Queens
11430	Queens
11432	Queens
11433	Queens
11434	Queens
11435	Queens
11436	Queens
11451	Queens
11691	Queens

ZIP	Borough
11692	Queens
11693	Queens
11694	Queens
11695	Queens
11697	Queens
99999	bogus ZIP

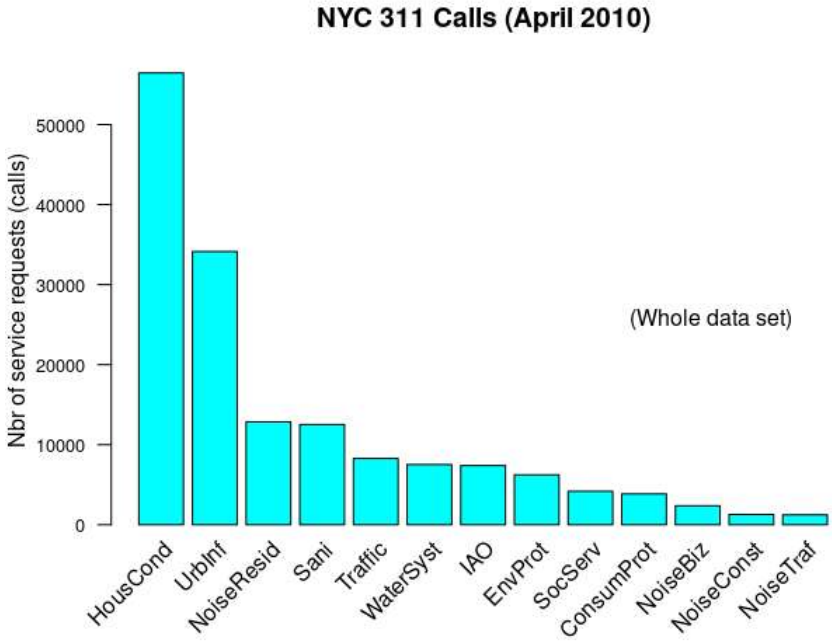
Appendix D: ZIP codes projection in PC1-2 after MCA and k-means/HC clustering (April 2014 data)



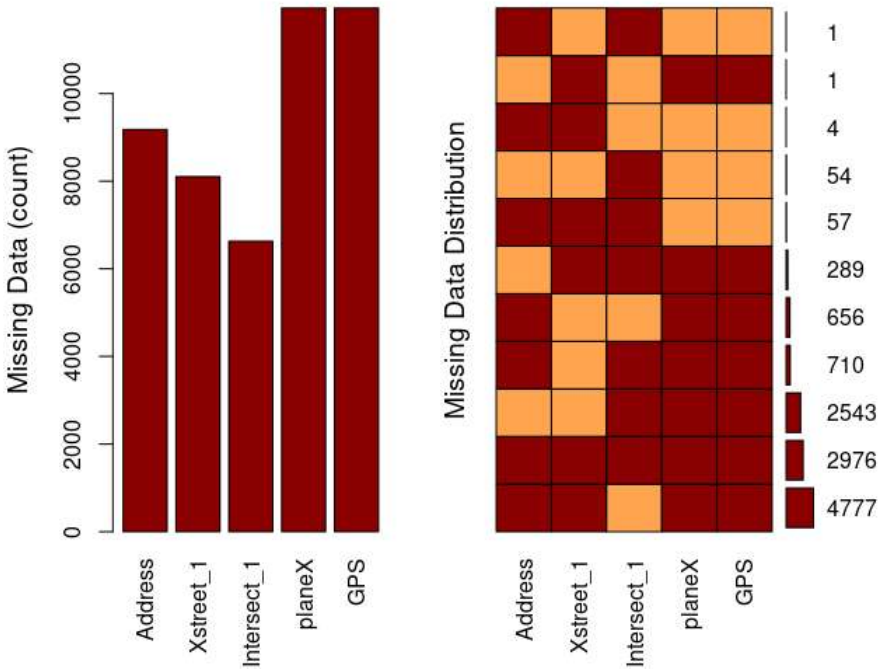
Appendix E: Topological representation after MCA and clustering, without consolidation – April 2014 NYC SRCs+crime data



Appendix F: Analytical results summary for the period April 2010




```
obs w/ missing ZIPs: 1206
obs w/ missing GPS coords: 19583
obs w/ missing ZIP and GPS coords: 11952
obs w/ missing ZIP, Address, GPS coords 9119
```



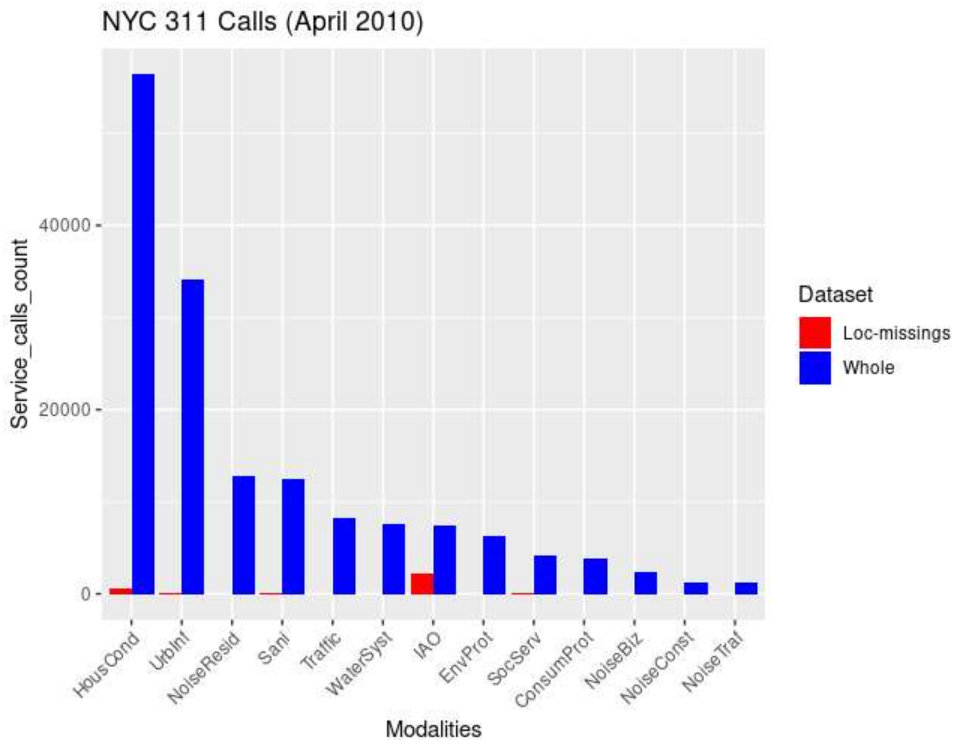
For all obs with missing ZIP, 2976 obs miss all geoloc info.

```
 Missings per variable:
    Variable Count
     Address  9181
   Xstreet_1  8104
 Intersect_1  6630
      planeX 11952
         GPS 11952

 Missings in combinations of variables:
 Combinations Count      Percent
    0:0:1:0:0    54  0.447464369
    0:0:1:1:1  2543 21.072257209
    0:1:0:1:1     1  0.008286377
    0:1:1:1:1   289  2.394763010
    1:0:0:1:1   656  5.435863441
    1:0:1:0:0     1  0.008286377
    1:0:1:1:1   710  5.883327809
    1:1:0:0:0     4  0.033145509
    1:1:0:1:1  4777 39.584023865
    1:1:1:0:0    57  0.472323500
    1:1:1:1:1  2976 24.660258535
```

modalities of “all missing” categories

```
IAO  HousCond  SocServ  UrbInf  Sani ConsumProt  WaterSyst  NoiseTraf  Traffic  EnvProt  NoiseBiz
2253    583       69      45     19      3          2          1          1      0         0
NoiseConst NoiseResid
    0          0
```

2976 obs (24% of all obs missing a ZIP code) have no geolocation onformation. Compare the modality distribution of those 2976 complaints (frequency wise) with those of the whole data set.



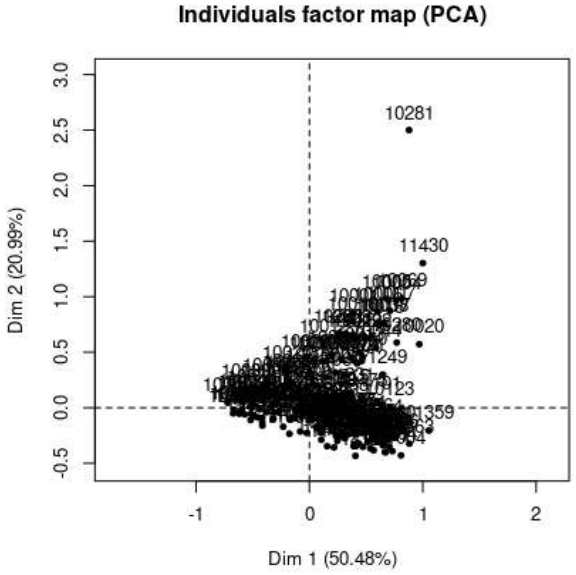
```
Fraction of modality 'SocServ':     1.65 %
Fraction of modality 'IAO':         30.44 %
Fraction of modality 'HousCond':    1.03 %
```

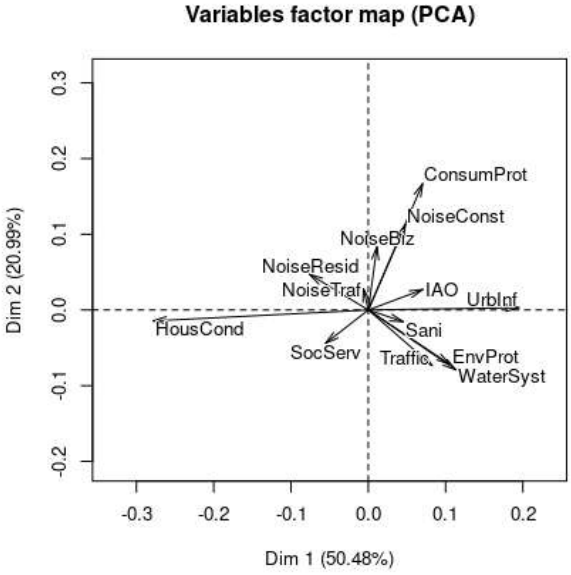
The Chi square test of independence of ZIP code observations and SRCs leads to a clear rejection of H_0.

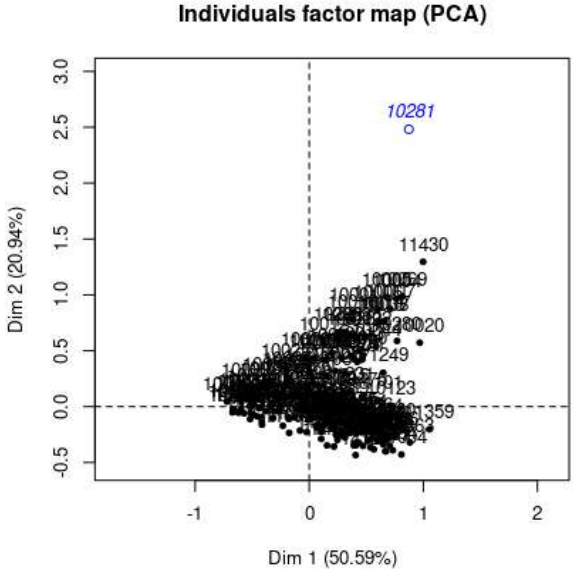
Nbr of cells with count <5: 453 . The contribution of low count cells to the computation of the Chi square statistic is negligible. Furthermore those cells do not contribute to a significant association between categorical variable sin the 2 way table of association. (Computed with a targeted Chi square test, we accept H_0)

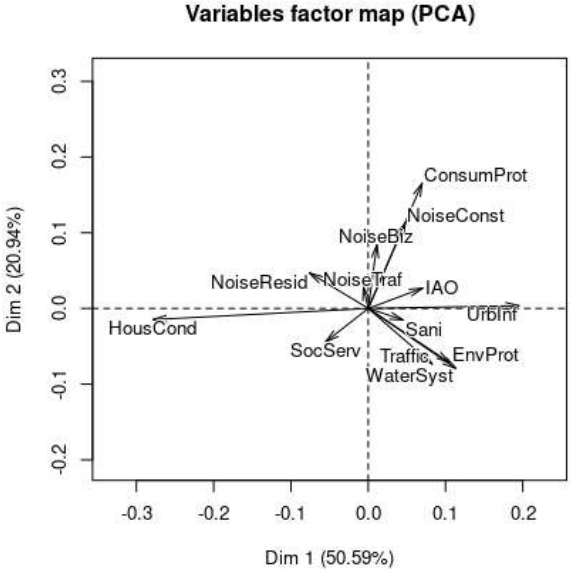
PCA on SRC for April 2010

Scatter plot of scores and variables' representation in PC1-2, PC1-3 and PC2-3
The effect of ZIP code "10281" is negligible. It is not a true outlier in this context of the April 2010 data set.









```
Eigenvalue percentage of variance cumulative percentage of variance
      (calculated considering "10281" as supplementary individual.)
comp 1  1.695898e-01           5.059007e+01                           50.59007
comp 2  7.020614e-02           2.094308e+01                           71.53315
comp 3  2.310180e-02           6.891460e+00                           78.42461
comp 4  1.481216e-02           4.418593e+00                           82.84321
comp 5  1.133313e-02           3.380768e+00                           86.22397
comp 6  9.811926e-03           2.926980e+00                           89.15095
comp 7  9.478464e-03           2.827506e+00                           91.97846
comp 8  8.528048e-03           2.543989e+00                           94.52245
comp 9  6.751816e-03           2.014124e+00                           96.53657
comp 10 6.534186e-03           1.949203e+00                           98.48578
comp 11 2.860668e-03           8.533615e-01                           99.33914
comp 12 2.215369e-03           6.608632e-01                          100.00000
comp 13 7.034108e-33           2.098334e-30                          100.00000
```

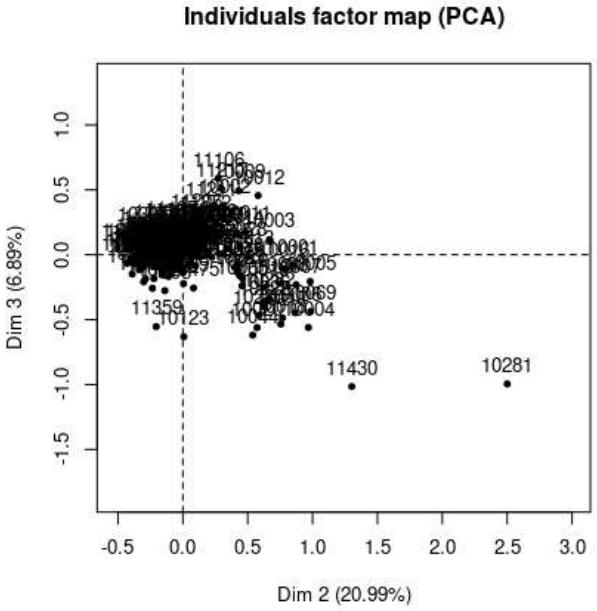
3 significant dimensions detected with criterion set at 72% of cumulated overall variance representation.

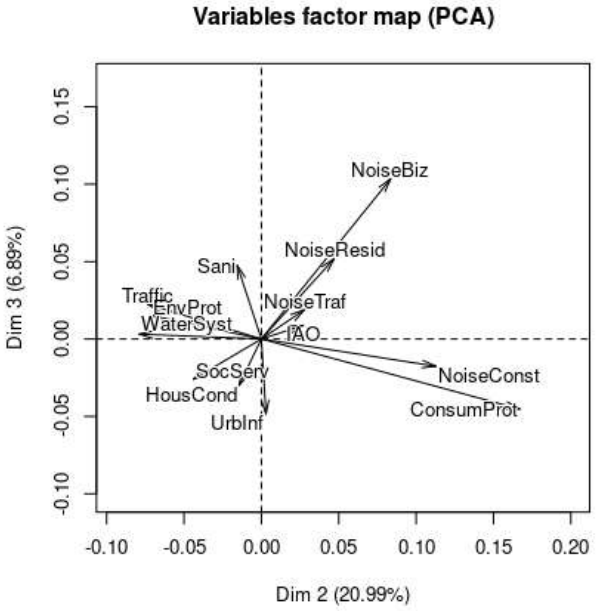
Correlation of variables with PCs:

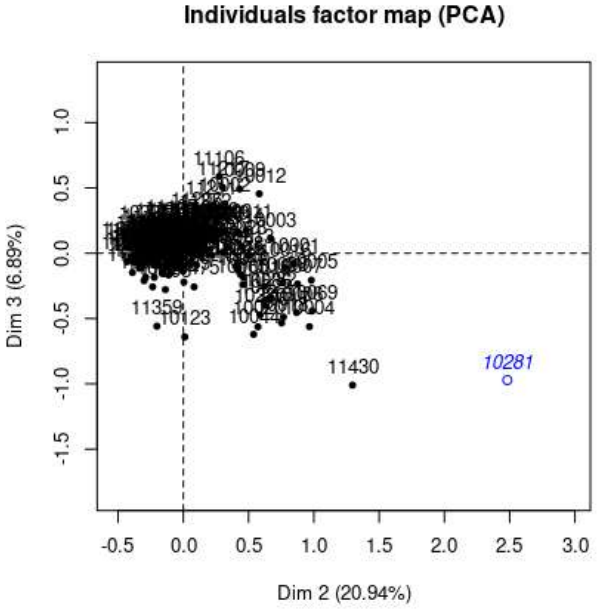
```
                Dim.1       Dim.2       Dim.3
HousCond   -0.98256559 -0.05182583 -0.10610209
IAO         0.71578939  0.27375519  0.08865394
NoiseResid -0.57383121  0.35411451  0.39149768
ConsumProt  0.35721157  0.84687174 -0.22758567
Traffic     0.56976394 -0.50522920  0.15725345
UrbInf      0.92581598  0.01690264 -0.23331906
NoiseBiz    0.07689977  0.57594401  0.70690277
WaterSyst   0.69485582 -0.48567646  0.02328010
NoiseConst  0.32326221  0.76196894 -0.12446854
Sani        0.42744793 -0.14454924  0.44858883
NoiseTraf  -0.10637282  0.47951506  0.32063047
EnvProt     0.71835778 -0.48830900  0.10476517
SocServ    -0.41506123 -0.33199723 -0.19106544
```

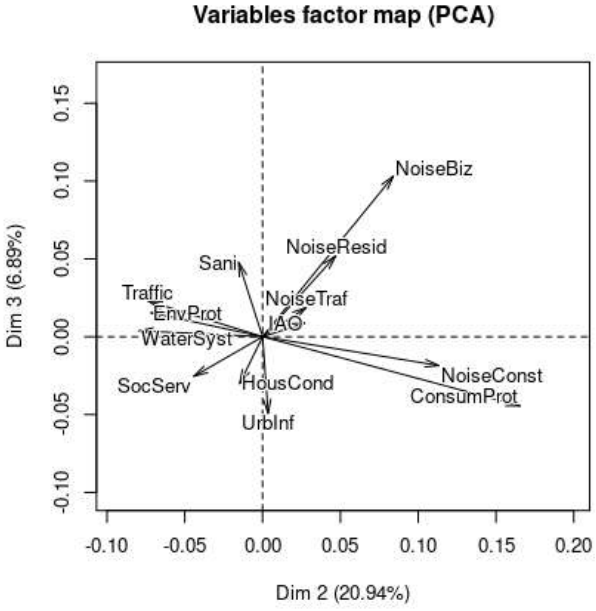
Contribution of varibales to the construction of each dimension

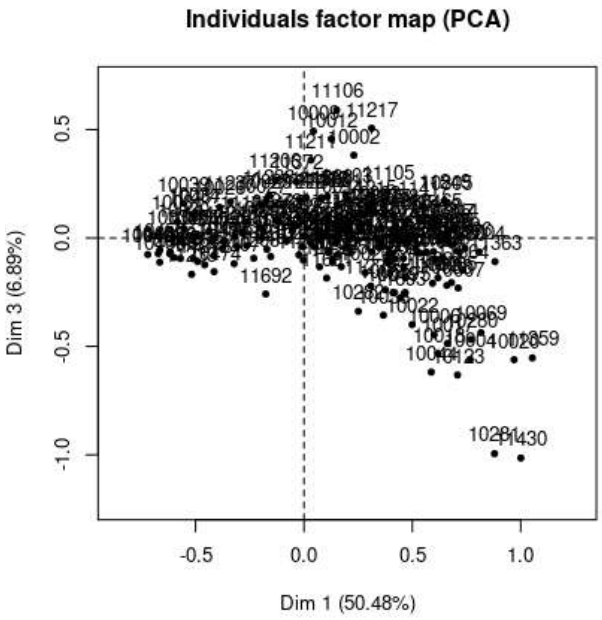
```
                 Dim.1       Dim.2       Dim.3
HousCond   45.68728963  0.30703616  3.91087716
IAO         2.94094250  1.03911926  0.33118155
NoiseResid  3.42038572  3.14643565 11.68743376
ConsumProt  2.87039255 38.97177644  8.55330787
Traffic     4.04335776  7.67986858  2.26103163
UrbInf     22.43838711  0.01806658 10.46156662
NoiseBiz    0.07427798 10.06456819 46.07691026
WaterSyst   7.57090740  8.93464491  0.06238529
NoiseConst  1.36301203 18.29317180  1.48341624
Sani        1.21658732  0.33607260  9.83620974
NoiseTraf   0.02272038  1.11527876  1.51536718
EnvProt     6.55367735  7.31505437  1.02327253
SocServ     1.79806227  2.77890669  2.79704018
```

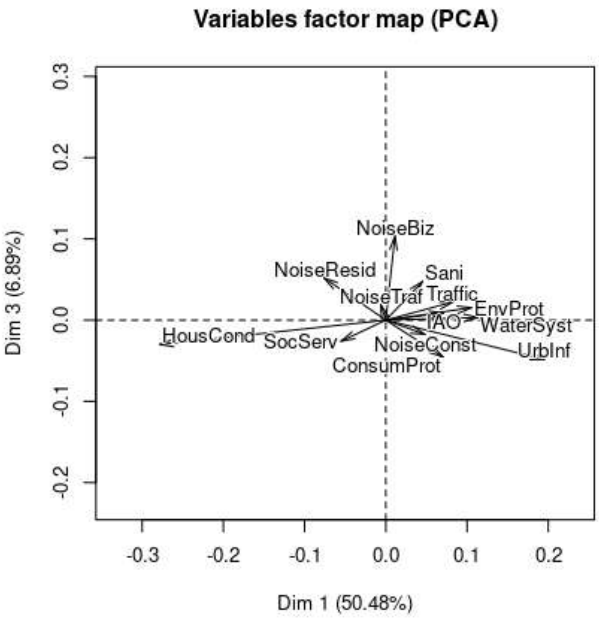

Individuals factor map (PCA)
Dim 3 (6.89%)
Dim 2 (20.99%)
11430
10281
11359
10123

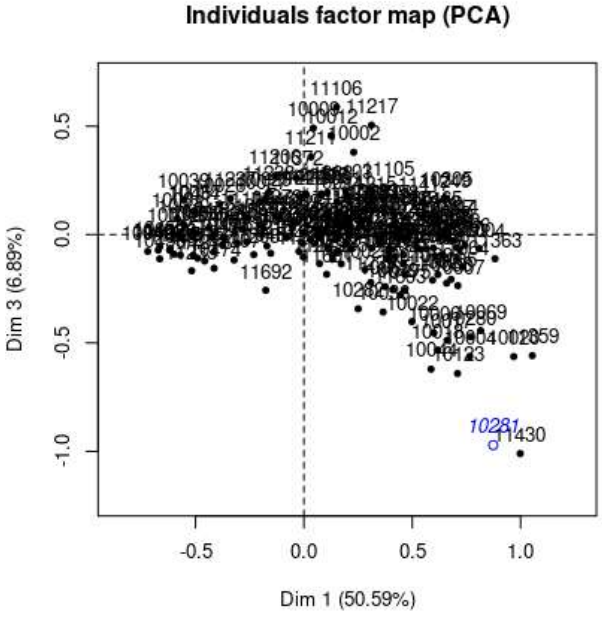

Variables factor map (PCA)
Dim 3 (6.89%)
Dim 2 (20.99%)
NoiseBiz
NoiseResid
Sani
Traffic
EnvProt
WaterSyst
NoiseTraf
IAO
SocServ
HousCond
UrbInf
NoiseConst
ConsumProt

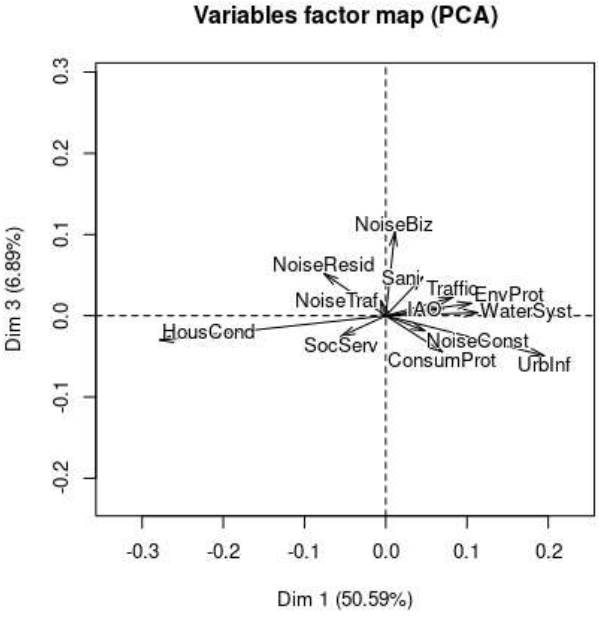

Individuals factor map (PCA)
Dim 3 (6.89%)
Dim 2 (20.94%)
11430
10281
11359
10123

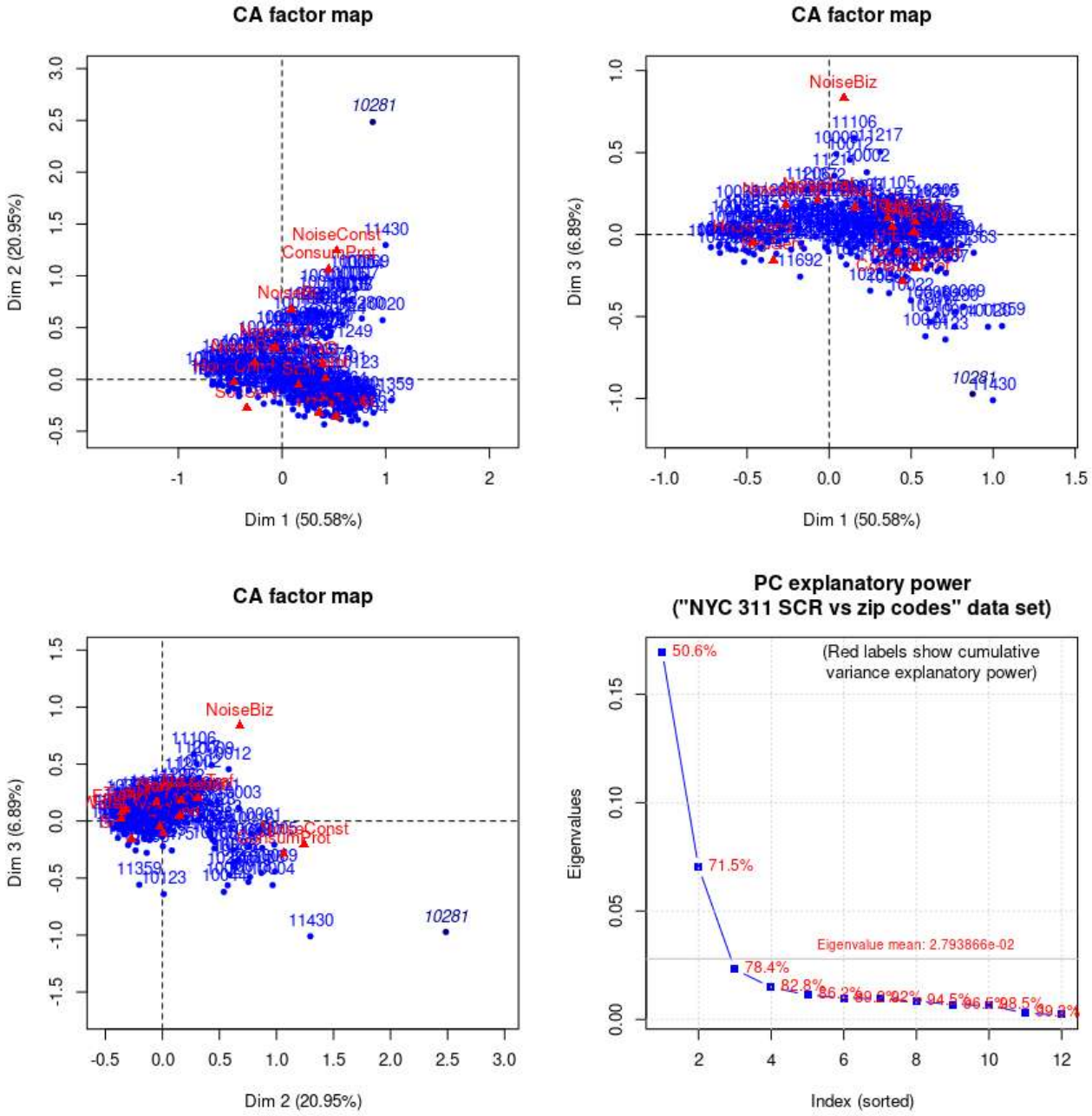

Variables factor map (PCA)
Dim 3 (6.89%)
Dim 2 (20.94%)
NoiseBiz
NoiseResid
Sani
Traffic
EnvProt
WaterSyst
NoiseTraf
IAO
SocServ
HousCond
UrbInf
NoiseConst
ConsumProt


Individuals factor map (PCA)
Dim 3 (6.89%)
Dim 1 (50.48%)
11106
11692
10281
11430


Variables factor map (PCA)
Dim 3 (6.89%)
Dim 1 (50.48%)
NoiseBiz
NoiseResid
Sani
NoiseTraf
Traffic
EnvProt
IAO
WaterSyst
HousCond
SocServ
NoiseConst
UrbInf
ConsumProt


Individuals factor map (PCA)
Dim 3 (6.89%)
Dim 1 (50.59%)
11106
11692
10281
11430


Variables factor map (PCA)
Dim 3 (6.89%)
Dim 1 (50.59%)
NoiseBiz
NoiseResid
Sani
Traffic
NoiseTraf
EnvProt
IAO
WaterSyst
HousCond
SocServ
NoiseConst
ConsumProt
UrbInf



Contributions to the construction of principal directions > 10%

	Dim 1	Dim 2	Dim 3
HousCond	45.68515128	0.30503717	3.912213
NoiseResid	3.42000942	3.14513175	11.690464
ConsumProt	2.87631691	39.00901070	8.556879
UrbInf	22.43720197	0.01779967	10.453181
NoiseBiz	0.07439473	10.05350513	46.086619
NoiseConst	1.36380297	18.27895121	1.477422

Inertia explanatory power for all dimensions and for the significant dimensions

	iep_alldim	iep_sigdim
HousCond	23.9	29.9
IAO	2.9	2.2
NoiseResid	5.3	4.1
ConsumProt	11.4	13.0
Traffic	6.3	4.9
UrbInf	13.2	15.4
NoiseBiz	6.4	6.8
WaterSyst	7.9	7.3
NoiseConst	6.6	5.9
Sani	3.4	1.7
NoiseTraf	1.0	0.4
EnvProt	6.4	6.3
SocServ	5.3	2.1

Quality of representation of col profiles with biggest contrib to PC formation

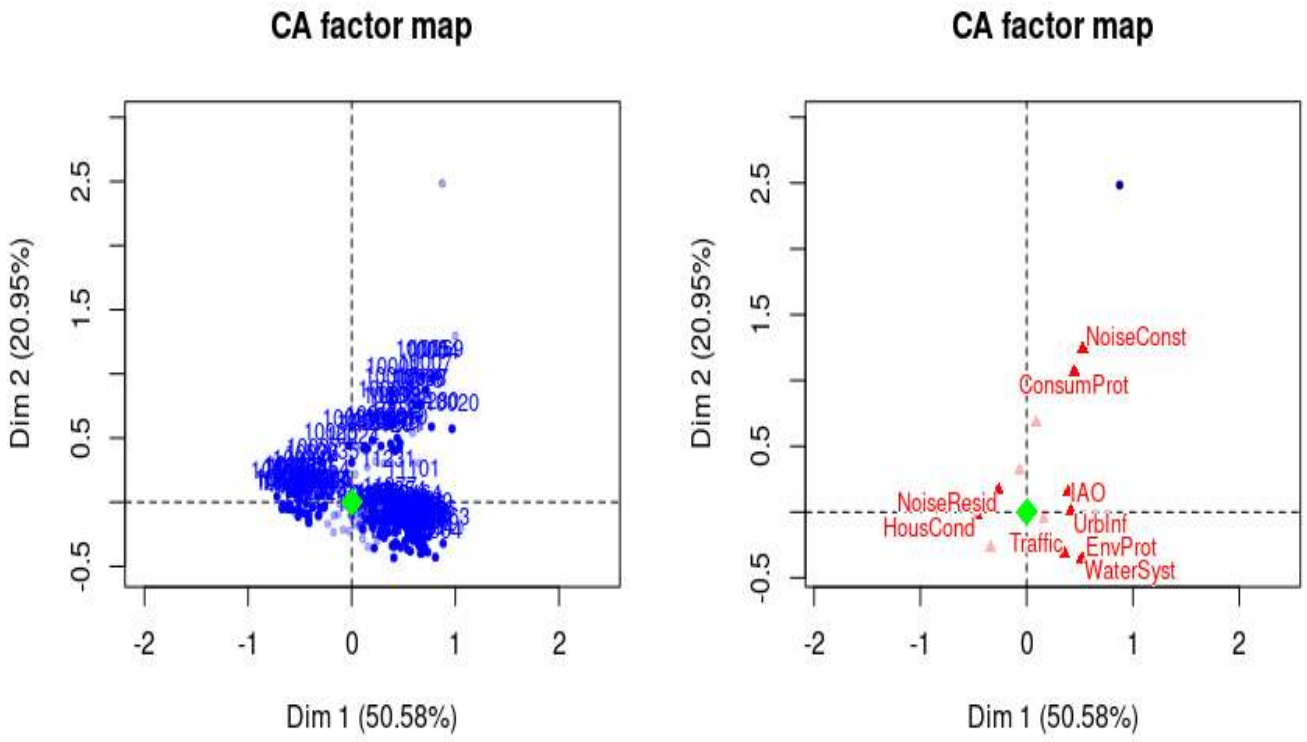
	Dim 1	Dim 2	Dim 3
HousCond	0.965444766	0.0026698605	0.01126289
NoiseResid	0.329239118	0.1254025611	0.15331723
ConsumProt	0.127700459	0.7173053924	0.05175427
UrbInf	0.857138637	0.0002816293	0.05440088
NoiseBiz	0.005923206	0.3315245110	0.49987871
NoiseConst	0.104565032	0.5804563311	0.01543172

Most important individual contributors to the construction of PCs

	Dim 1	Dim 2	Dim 3	
10001	0.4	5.0	0.1	
10002	0.3	1.4	5.8	
10003	0.5	5.1	0.5	
10007	0.6	2.1	0.5	
10009	0.0	1.8	6.9	**Dim3** Alphabet City, Downtown Manhattan
10011	0.2	2.3	0.9	
10012	0.0	2.2	4.1	
10013	0.7	2.3	0.0	
10016	0.6	5.7	0.4	**Dim2** Murray Hill in South Central Manhattan
10017	0.8	2.8	3.4	
10018	0.6	2.3	3.5	
10019	0.7	3.9	2.0	
10022	0.7	2.9	3.5	
10031	2.1	0.1	0.3	
10032	2.3	0.1	0.0	
10036	0.5	3.2	3.2	
10040	3.0	0.0	0.3	
10314	2.3	0.9	0.1	
10452	3.5	0.0	0.2	
10453	3.9	0.1	0.3	
10457	2.4	0.1	0.5	
10458	3.7	0.0	0.8	
10460	2.2	0.0	0.2	

```
10467    2.2    0.0    0.7
11106    0.1    0.5    6.3
11206    0.1    0.1    2.5
11211    0.0    0.7    6.0
11217    0.3    0.6    5.2
11225    2.1    0.0    0.4
11226    4.0    0.1    0.4    Dim1  Center of Brooklyn, near Prospect Park
11430    0.5    1.9    3.5
```

Results for CA:



Left row profiles with cos² > 0.6 – Right col profiles with cos² > 0.4

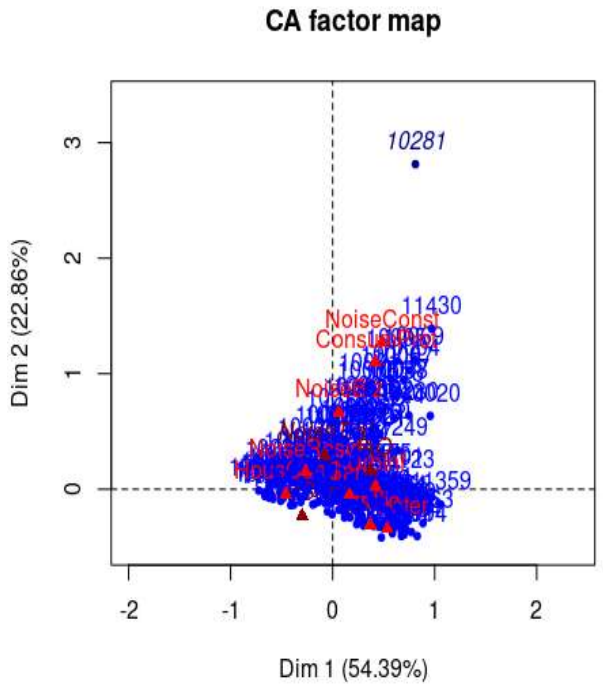
After feature selection and extraction:

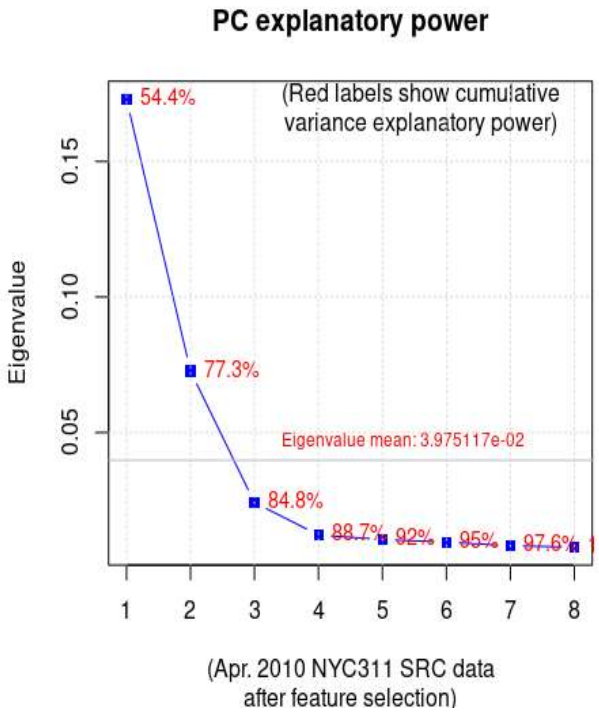
"EnvProt" and "WaterSyst" are strongly correlated in PC1-2-3
"IAO", "NoiseTraf" and "SocServ" are weakly correlated with PCs and can be dispensed with.

```
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.172970215              54.391548                           54.39155
dim 2 0.072692304              22.858542                           77.25009
dim 3 0.024123041               7.585639                           84.83573
dim 4 0.012197252               3.835501                           88.67123
dim 5 0.010517256               3.307216                           91.97845
dim 6 0.009583801               3.013685                           94.99213
dim 7 0.008195555               2.577143                           97.56927
```

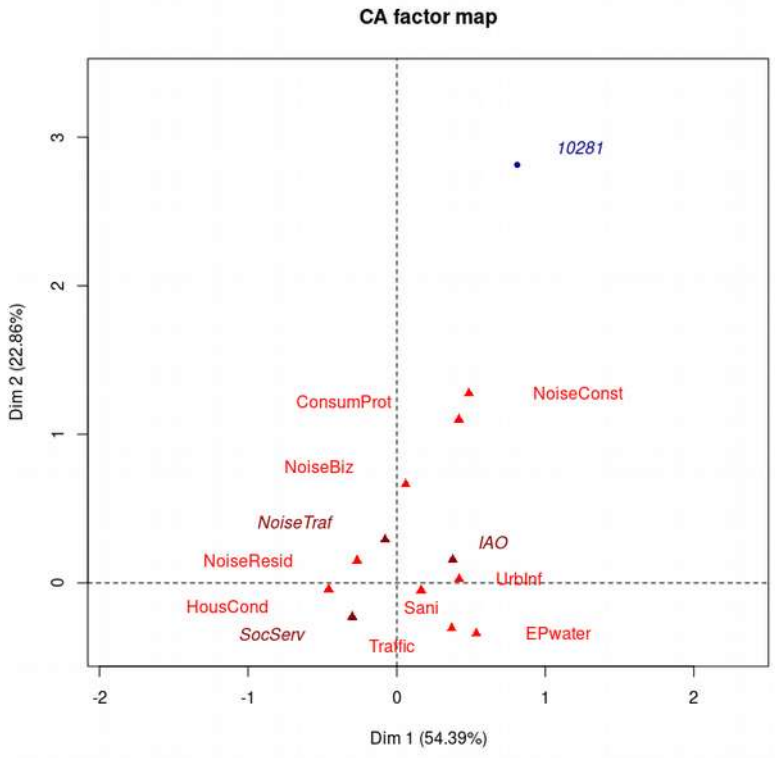
```
dim 8 0.007729936              2.430726                    100.00000
```

→ 2 significant dimensions only





The chi square test of independence between the two variables (ZIPs + SRCs) is equal to 45880.48 (p-value = 0).



Columns

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
HousCond	83.937	-0.458	46.906	0.967	-0.042	0.934	0.008	-0.060	5.680	0.016
NoiseResid	18.725	-0.266	3.649	0.337	0.151	2.787	0.108	0.185	12.607	0.162
ConsumProt	41.344	0.418	2.631	0.110	1.100	43.271	0.761	-0.225	5.438	0.032
Traffic	22.625	0.370	4.544	0.347	-0.306	7.395	0.238	0.095	2.123	0.023
UrbInf	47.782	0.421	23.903	0.865	0.026	0.218	0.003	-0.105	10.767	0.054
NoiseBiz	23.032	0.060	0.035	0.003	0.664	9.955	0.314	0.860	50.349	0.527
EPwater	44.201	0.536	15.757	0.617	-0.341	15.220	0.250	0.040	0.637	0.003
NoiseConst	24.135	0.486	1.218	0.087	1.275	19.940	0.601	-0.180	1.203	0.012
Sani	12.228	0.165	1.358	0.192	-0.049	0.281	0.017	0.177	11.195	0.221

Active variable are in red, while supplementary ones are in dark red. The blue point is the supplementary individual ear-marked as outlier.

Contributions (> 10%) to the construction of the 3 first PCs after feature extraction/selection:

	Dim 1	Dim 2	Dim 3
HousCond	46.90561416	0.9342768	5.6802232
NoiseResid	3.64862614	2.7871138	12.6072394
ConsumProt	2.63127856	43.2708473	5.4382810
UrbInf	23.90250351	0.2179053	10.7666437
NoiseBiz	0.03468261	9.9548084	50.3492163
EPwater	15.75720261	15.2196609	0.6372687
NoiseConst	1.21836939	19.9400153	1.2031320
Sani	1.35791871	0.2808020	11.1950591

Quality of representations within the former group (above):

	Dim 1	Dim 2	Dim 3
HousCond	0.966586741	0.008091120	0.016324585
NoiseResid	0.337044153	0.108200496	0.162419498
ConsumProt	0.110084673	0.760803565	0.031730901
UrbInf	0.865264711	0.003315054	0.054355898
NoiseBiz	0.002604616	0.314182419	0.527333437
EPwater	0.616625987	0.250302272	0.003477972
NoiseConst	0.087318834	0.600580947	0.012025495
Sani	0.192081416	0.016692789	0.220850659

Individual contributions for which at least one component is greater than 20% in PC1-2-3

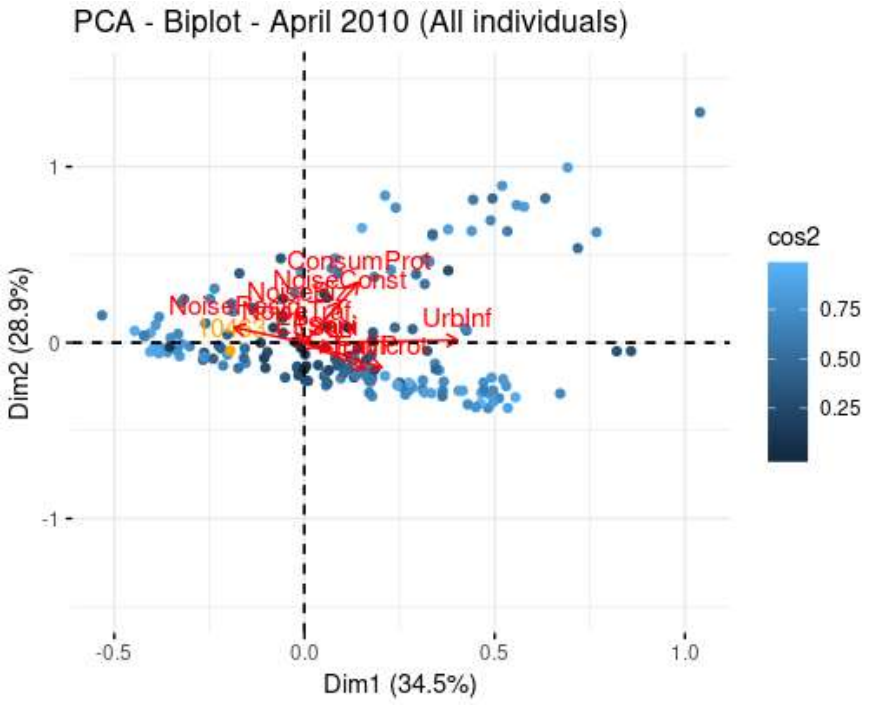
	Dim 1	Dim 2	Dim 3	
10001	0.4	5.3	0.0	
10002	0.3	1.3	6.1	
10003	0.4	5.4	0.7	
10007	0.5	2.3	0.3	
10009	0.0	1.7	7.6	**Dim3**
10011	0.1	2.4	1.3	
10012	0.0	2.1	4.8	
10013	0.7	2.4	0.0	
10016	0.4	6.2	0.2	**Dim2**
10017	0.8	3.2	3.0	
10018	0.6	2.5	3.1	
10019	0.6	4.1	1.9	
10022	0.7	3.2	2.9	
10031	2.3	0.0	0.1	

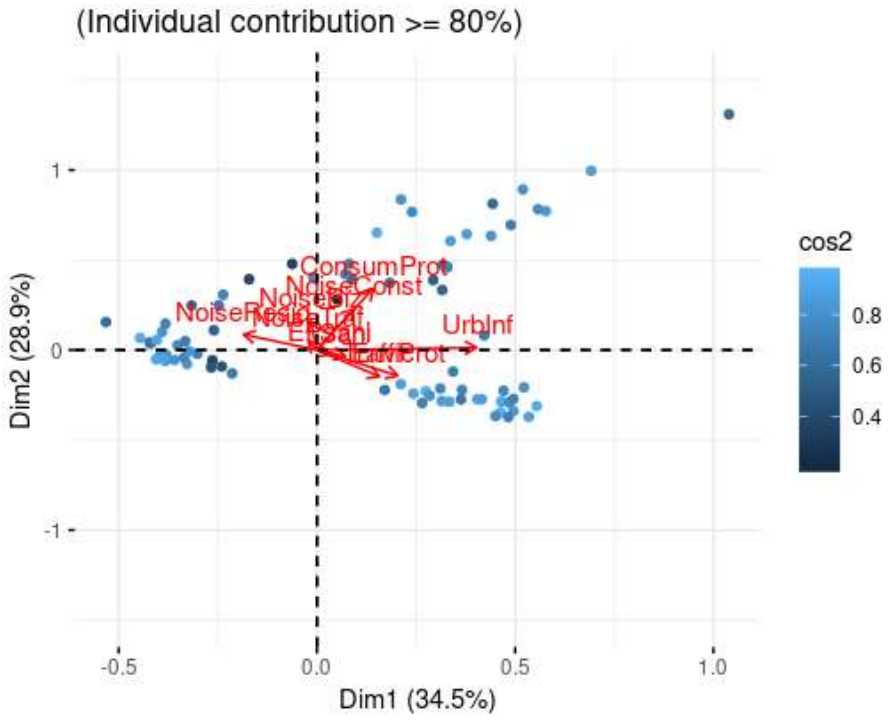
```
10032    2.4    0.0    0.1
10036    0.4    3.6    2.6
10040    3.1    0.0    0.7
10314    2.5    0.8    0.0
10452    3.4    0.0    0.2
10453    3.8    0.1    0.3
10457    2.5    0.1    0.6
10458    3.7    0.0    1.0
10460    2.1    0.0    0.1
11106    0.0    0.4    6.5
11206    0.1    0.1    2.8
11211    0.0    0.5    5.3
11217    0.2    0.6    5.5
11225    2.2    0.0    0.7
11226    4.0    0.1    0.6 Dim1
11372    0.0    0.2    2.2
11430    0.4    2.2    2.8
```

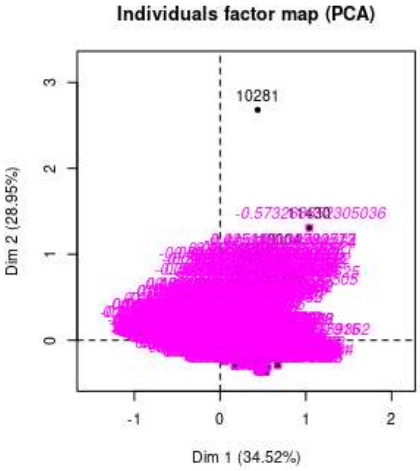
Not changed from before. Good sign.

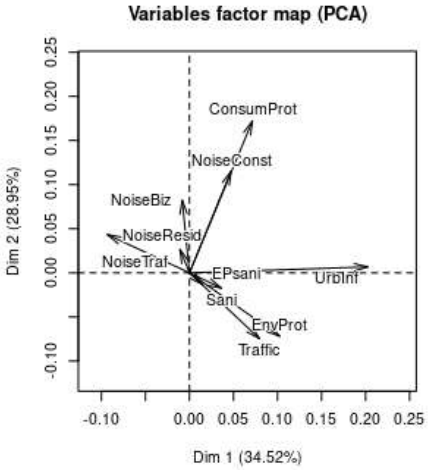
Inertia Explanatory Power for factors (i.e. variable's modalities)

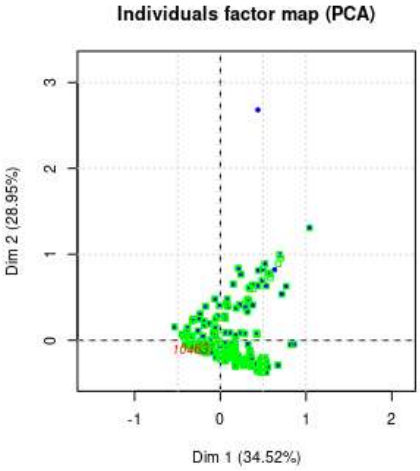
```
           iep_alldim iep_sigdim
HousCond         26.4       33.3
NoiseResid        5.9        3.4
ConsumProt       13.0       14.7
Traffic           7.1        5.4
UrbInf           15.0       16.9
NoiseBiz          7.2        3.0
EPwater          13.9       15.6
NoiseConst        7.6        6.8
Sani              3.8        1.0
```

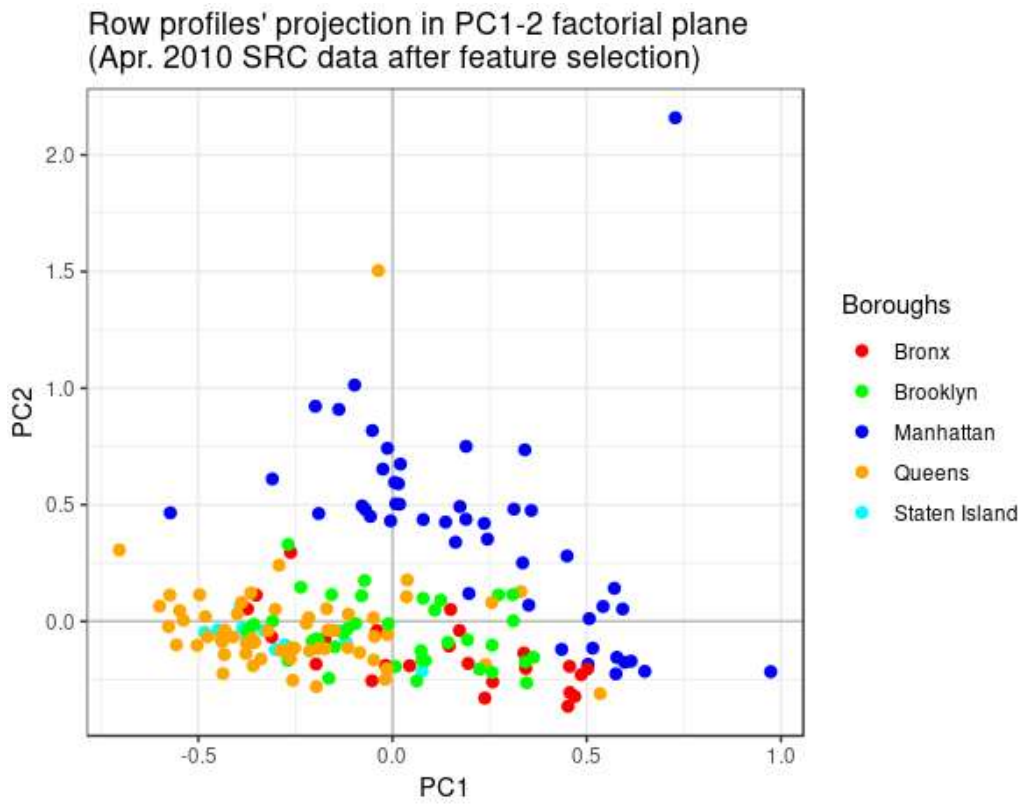



(Individual contribution >= 80%)
Dim2 (28.9%)
Dim1 (34.5%)
cos2
0.8
0.6
0.4
ConsumProt
NoiseConst
NoiseResid
UrbInf
-0.5
0.0
0.5
1.0
1
0
-1


Individuals factor map (PCA)
Dim 2 (28.95%)
Dim 1 (34.52%)
10281

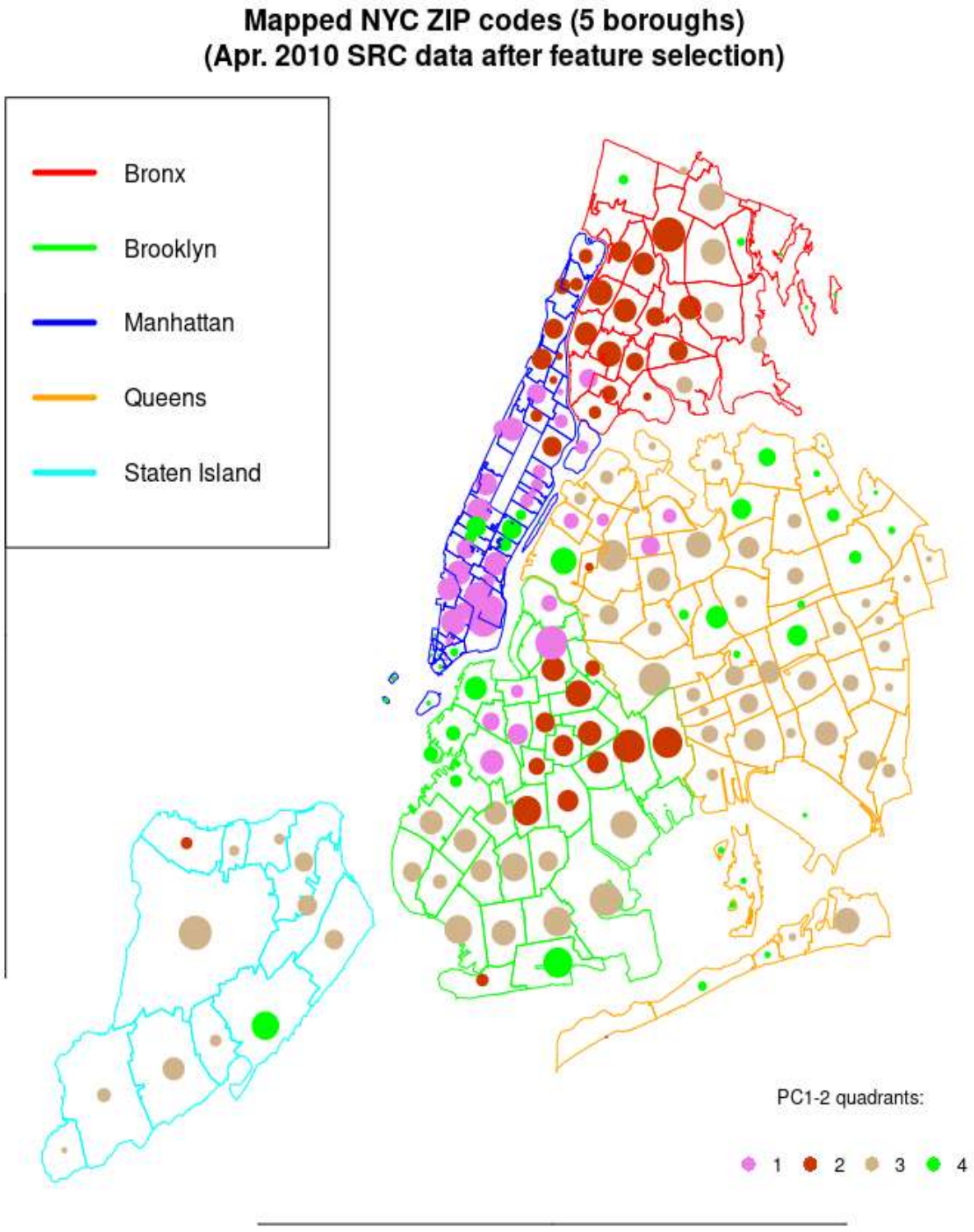

Variables factor map (PCA)
Dim 2 (28.95%)
Dim 1 (34.52%)
ConsumProt
NoiseConst
NoiseBiz
NoiseResid
NoiseTraf
EPsani
UrbInf
Sani
EnvProt
Traffic


Individuals factor map (PCA)
Dim 2 (28.95%)
Dim 1 (34.52%)

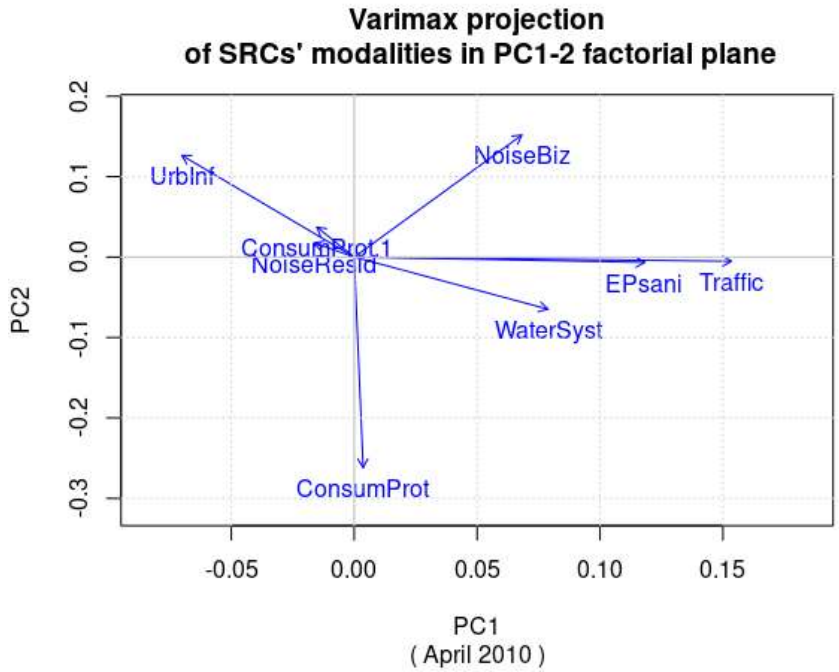


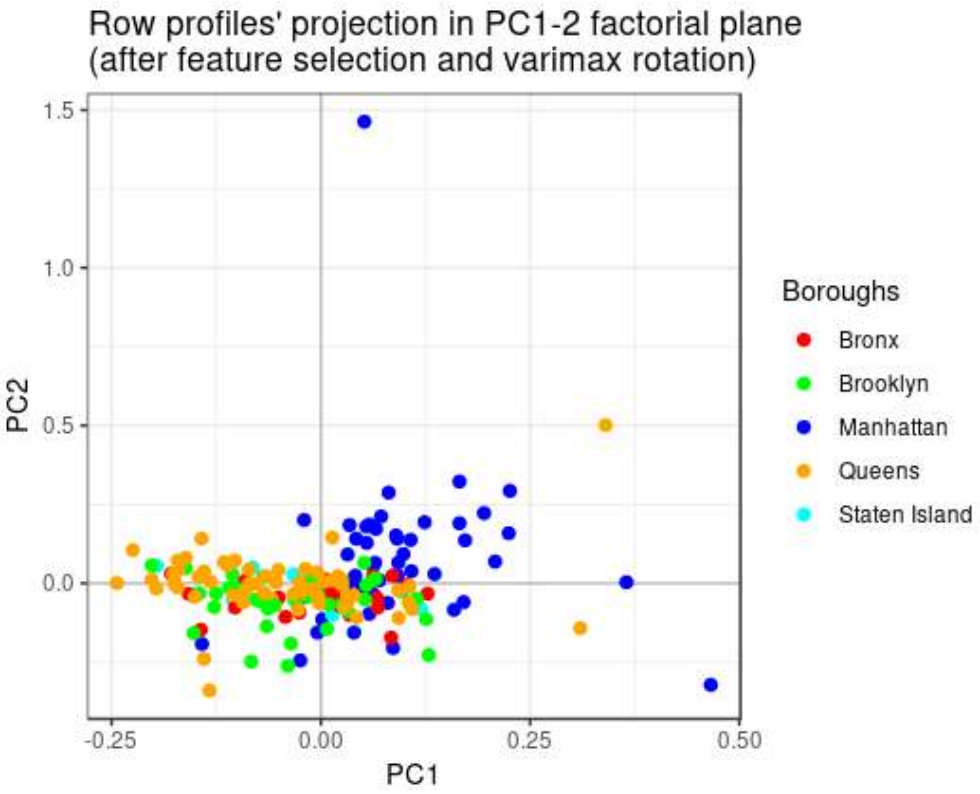
<u>Analysis of ZIP codes' (observations') contributions to inertia per borough</u>:

```
Borough:  Manhattan
 Number of ZIP codes:  47
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   42.1    30.1

Borough:  Staten Island
 Number of ZIP codes:  12
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
      7       5

Borough:  Bronx
 Number of ZIP codes:  24
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   10.4     7.9

Borough:  Queens
 Number of ZIP codes:  60
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   23.6    12.2

Borough:  Brooklyn
 Number of ZIP codes:  38
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   15.7     7.7
```

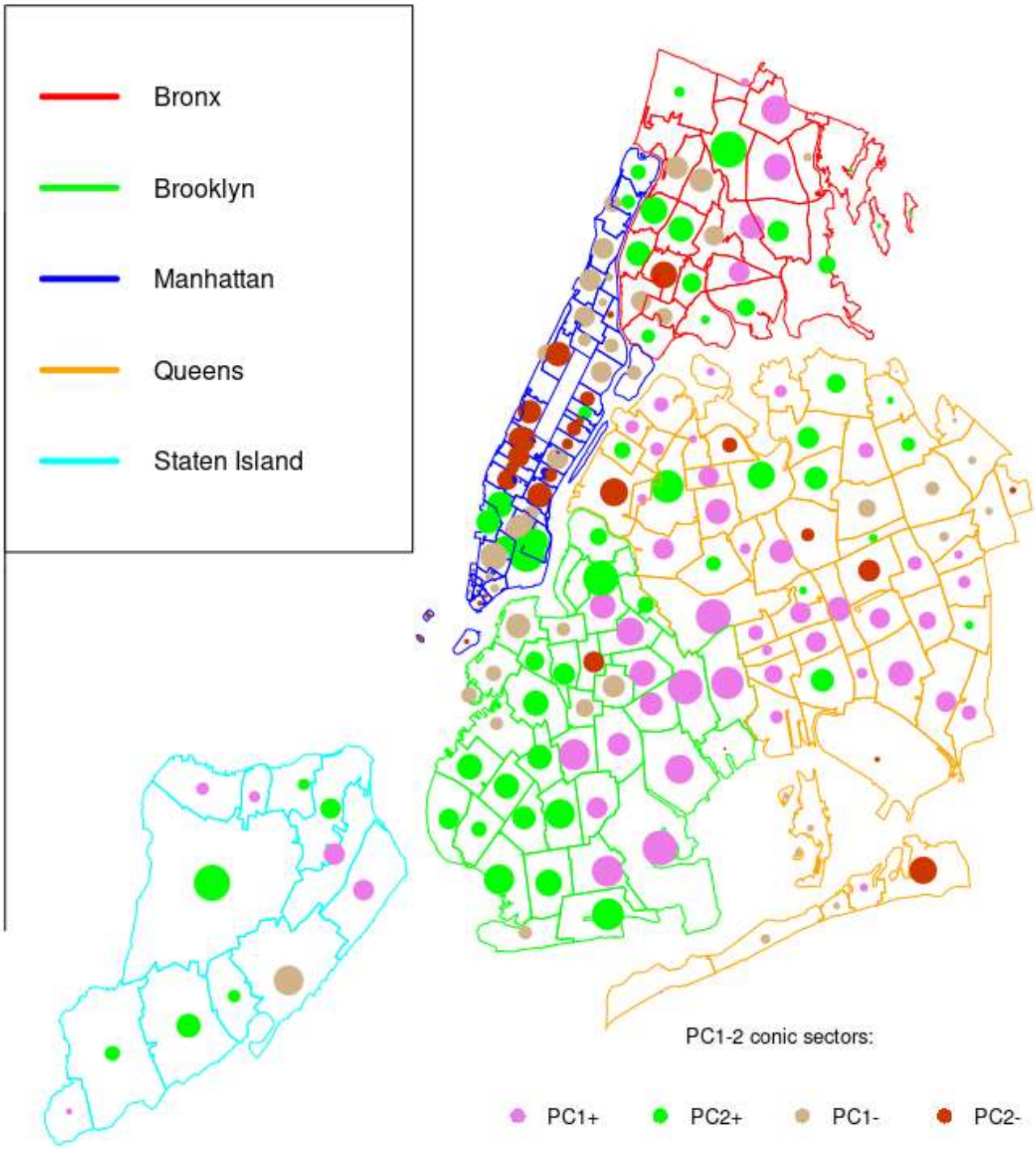

Mapped NYC ZIP codes (5 boroughs)
(Apr. 2010 SRC data after feature selection)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
PC1-2 quadrants:
1
2
3
4

Analysis of latent features following varimax treatment:





Mapped NYC ZIP codes (5 boroughs) - Apr. 2010
(after feature selection and varimax rotation)



MCA with crime data:

Binning the 17 categorical variables brings some change.

▪ Binning the 17 categorical variables brings some change when compared to reference year 2014, in particular in the distribution and frequency of "HousCond" related SRCs.

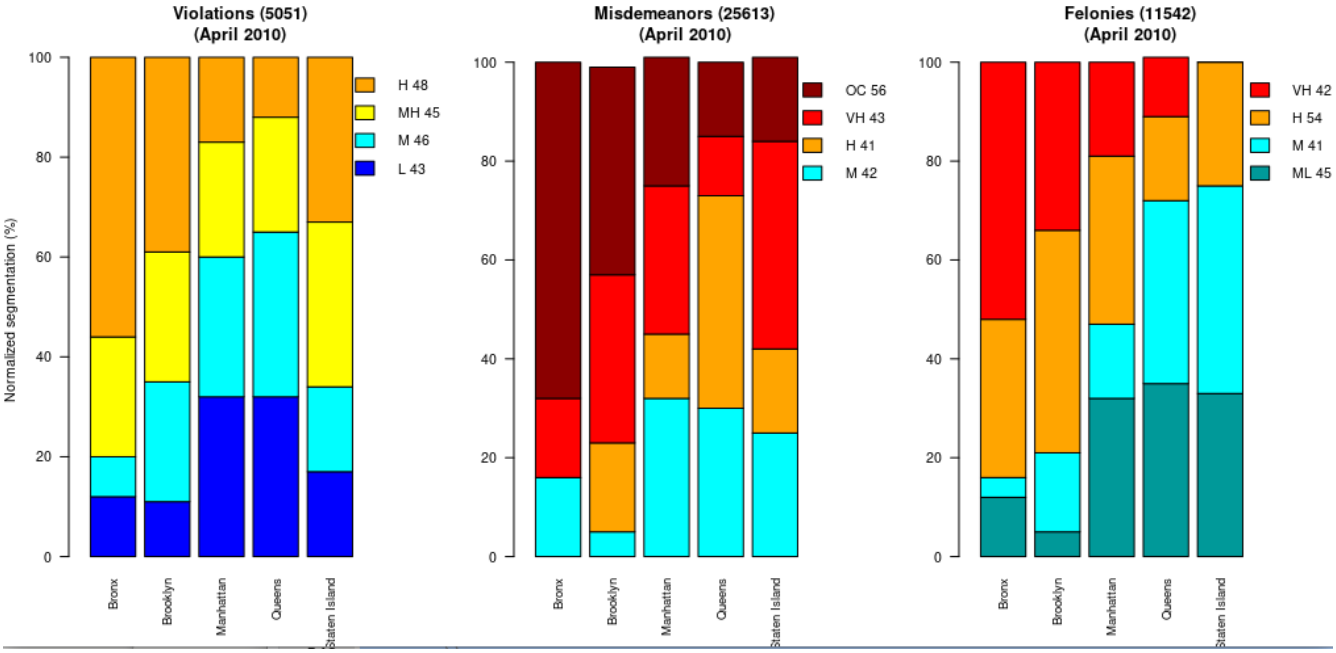
```
HousCond:
   below or equal to 80 SRCs - bin count: 47
   between 81 and 207 SRCs - bin count: 45
   between 208 and 406 SRCs- bin count: 45
   above 406 SRCs - bin count: 46
Sani:
   below or equal to 18 SRCs - bin count: 26
   between 19 and 31 SRCs - bin count: 14
   between 32 and 54 SRCs- bin count: 41
   above 54 SRCs - bin count: 102
NoiseResid:
   below or equal to 29 SRCs - bin count: 57
   between 30 and 61 SRCs - bin count: 40
   between 62 and 123 SRCs- bin count: 53
   above 123 SRCs - bin count: 33
NoiseConst:
   below or equal to 1 SRCs - bin count: 63
   between 2 and 5 SRCs - bin count: 63
   between 6 and 20 SRCs- bin count: 40
   above 20 SRCs - bin count: 17
NoiseBiz:
   below or equal to 3 SRCs - bin count: 60
   between 4 and 9 SRCs - bin count: 57
   between 10 and 27 SRCs- bin count: 42
   above 27 SRCs - bin count: 24
UrbInf:
   below or equal to 33 SRCs - bin count: 11
   between 34 and 55 SRCs - bin count: 6
   between 56 and 87 SRCs- bin count: 12
   above 87 SRCs - bin count: 154
Traffic:
   below or equal to 24 SRCs - bin count: 66
   between 25 and 54 SRCs - bin count: 58
   between 55 and 87 SRCs- bin count: 30
   above 87 SRCs - bin count: 29
NoiseTraf:
   below or equal to 4 SRCs - bin count: 89
   between 5 and 11 SRCs - bin count: 53
   between 12 and 23 SRCs- bin count: 35
   above 23 SRCs - bin count: 6
WaterSyst:
   below or equal to 19 SRCs - bin count: 41
   between 20 and 29 SRCs - bin count: 40
   between 30 and 44 SRCs- bin count: 36
   above 44 SRCs - bin count: 66
ConsumProt:
   below or equal to 5 SRCs - bin count: 44
   between 6 and 13 SRCs - bin count: 47
   between 14 and 23 SRCs- bin count: 39
   above 23 SRCs - bin count: 53
SocServ:
   below or equal to 2 SRCs - bin count: 33
   between 3 and 6 SRCs - bin count: 27
```

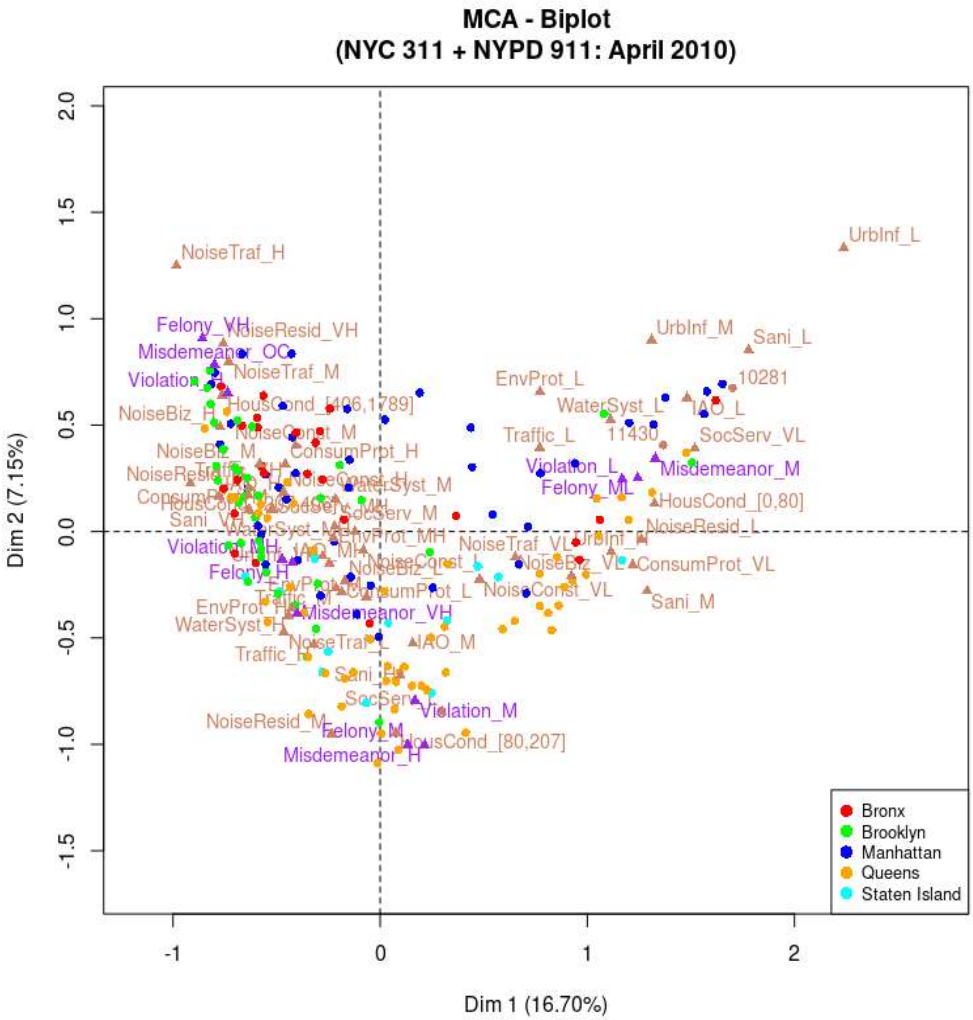
```
   between 7 and 11 SRCs- bin count: 28
   above 11 SRCs - bin count: 95
IAO:
   below or equal to 14 SRCs - bin count: 35
   between 15 and 23 SRCs - bin count: 46
   between 24 and 33 SRCs- bin count: 42
   above 33 SRCs - bin count: 60
EnvProt:
   below or equal to 16 SRCs - bin count: 51
   between 17 and 26 SRCs - bin count: 29
   between 27 and 41 SRCs- bin count: 44
   above 41 SRCs - bin count: 59
Violation:
   below or equal to 7 SRCs - bin count: 44
   between 8 and 20 SRCs - bin count: 46
   between 21 and 38 SRCs- bin count: 45
   above 38 SRCs - bin count: 48
Misdemeanor:
   below or equal to 33 SRCs - bin count: 43
   between 34 and 87 SRCs - bin count: 41
   between 88 and 178 SRCs- bin count: 43
   above 178 SRCs - bin count: 56
Felony:
   below or equal to 17 SRCs - bin count: 46
   between 18 and 45 SRCs - bin count: 41
   between 46 and 91 SRCs- bin count: 54
   above 91 SRCs - bin count: 42
```

	Violations			
Borough	L	M	MH	H
Bronx	3	2	6	14
Brooklyn	4	9	10	15
Manhattan	15	13	11	8
Queens	19	20	14	7
Staten Island	2	2	4	4

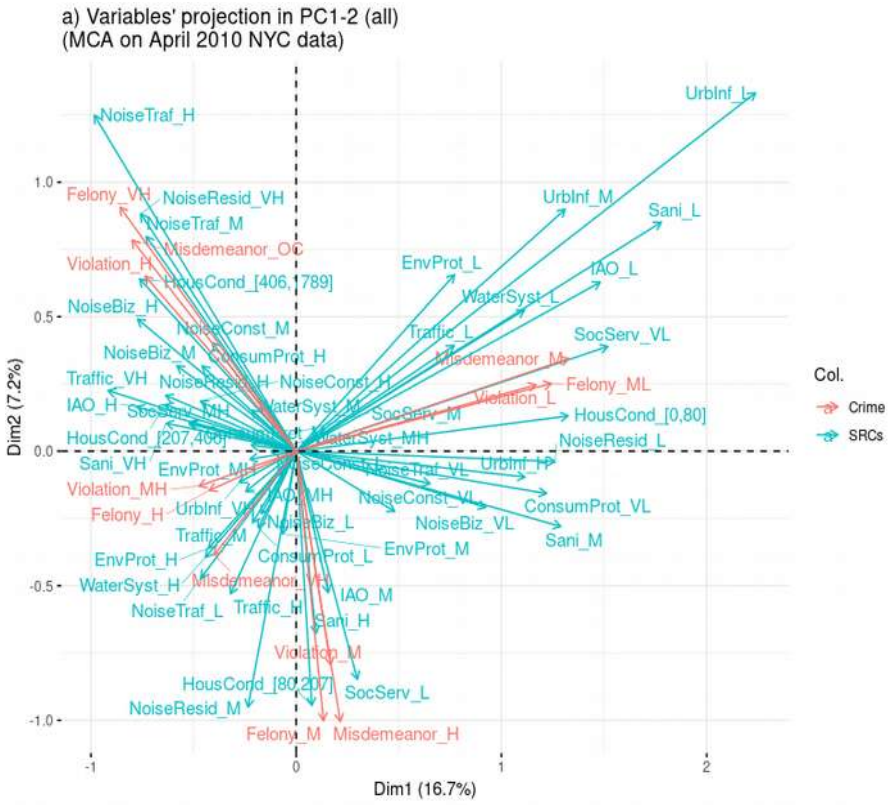
	Misdemeanor			
Borough	M	H	VH	OC
Bronx	4	0	4	17
Brooklyn	2	7	13	16
Manhattan	15	6	14	12
Queens	18	26	7	9
Staten Island	3	2	5	2

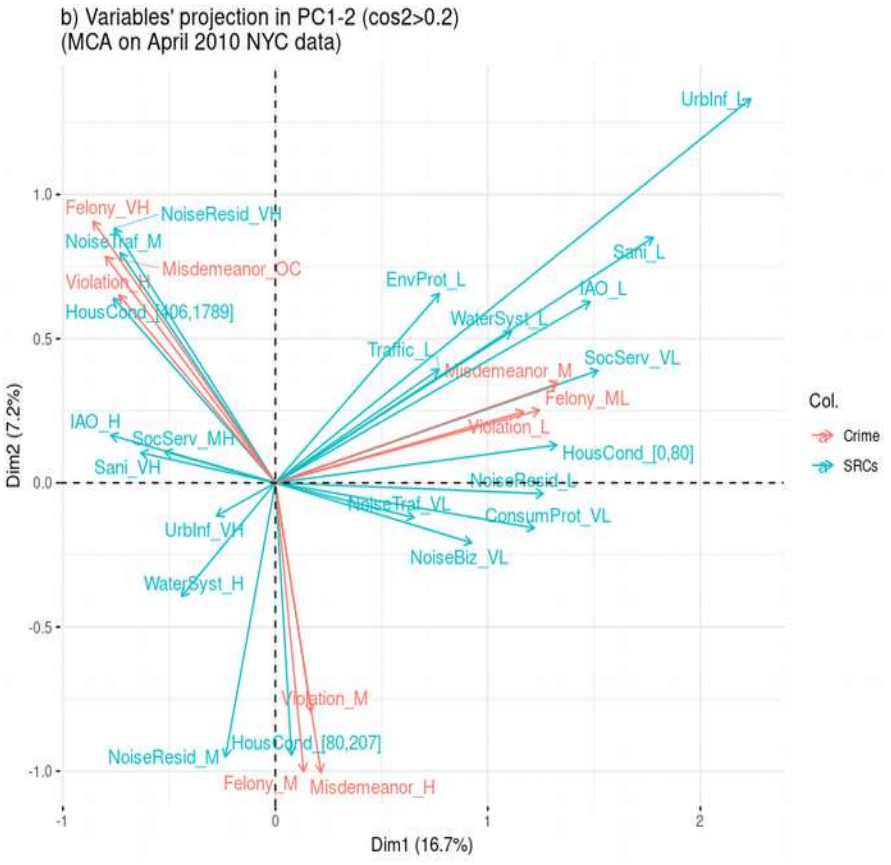
	Felony			
Borough	ML	M	H	VH
Bronx	3	1	8	13
Brooklyn	2	6	17	13
Manhattan	15	7	16	9
Queens	21	22	10	7
Staten Island	4	5	3	0

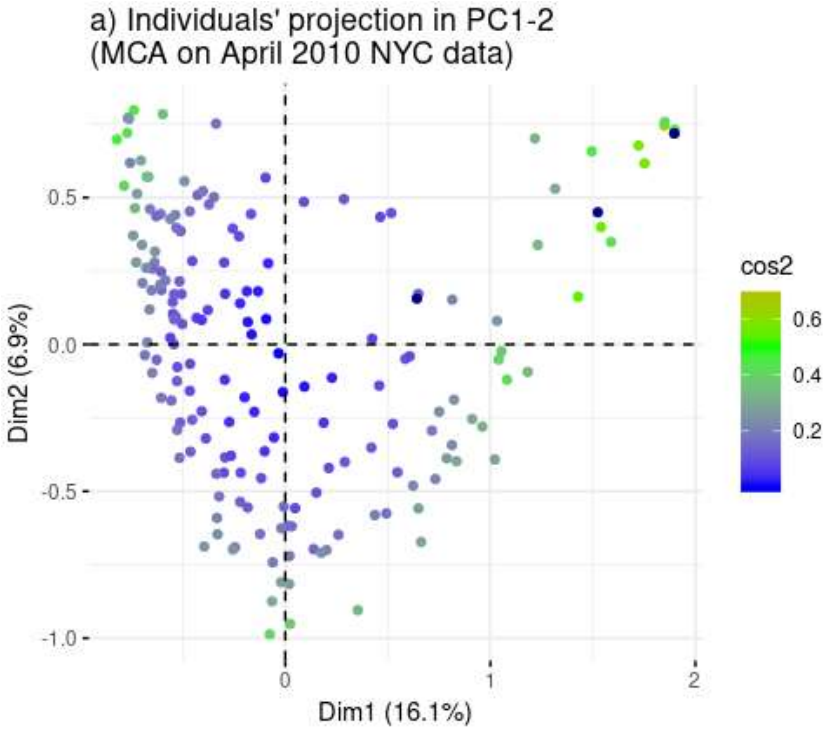


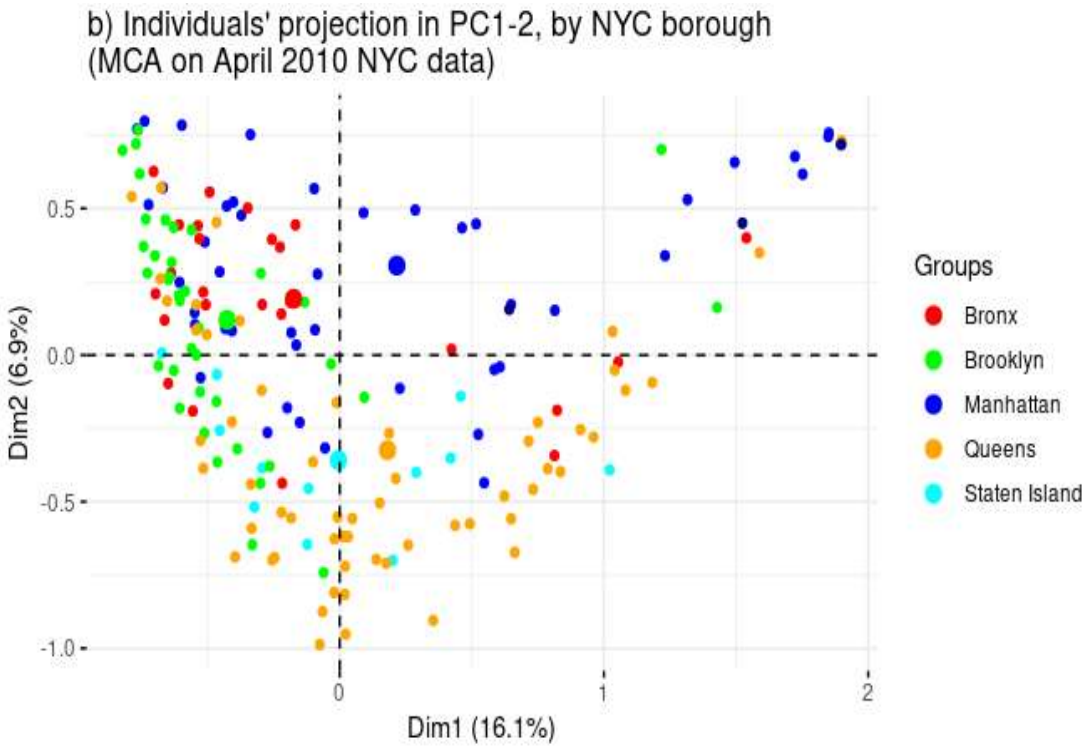


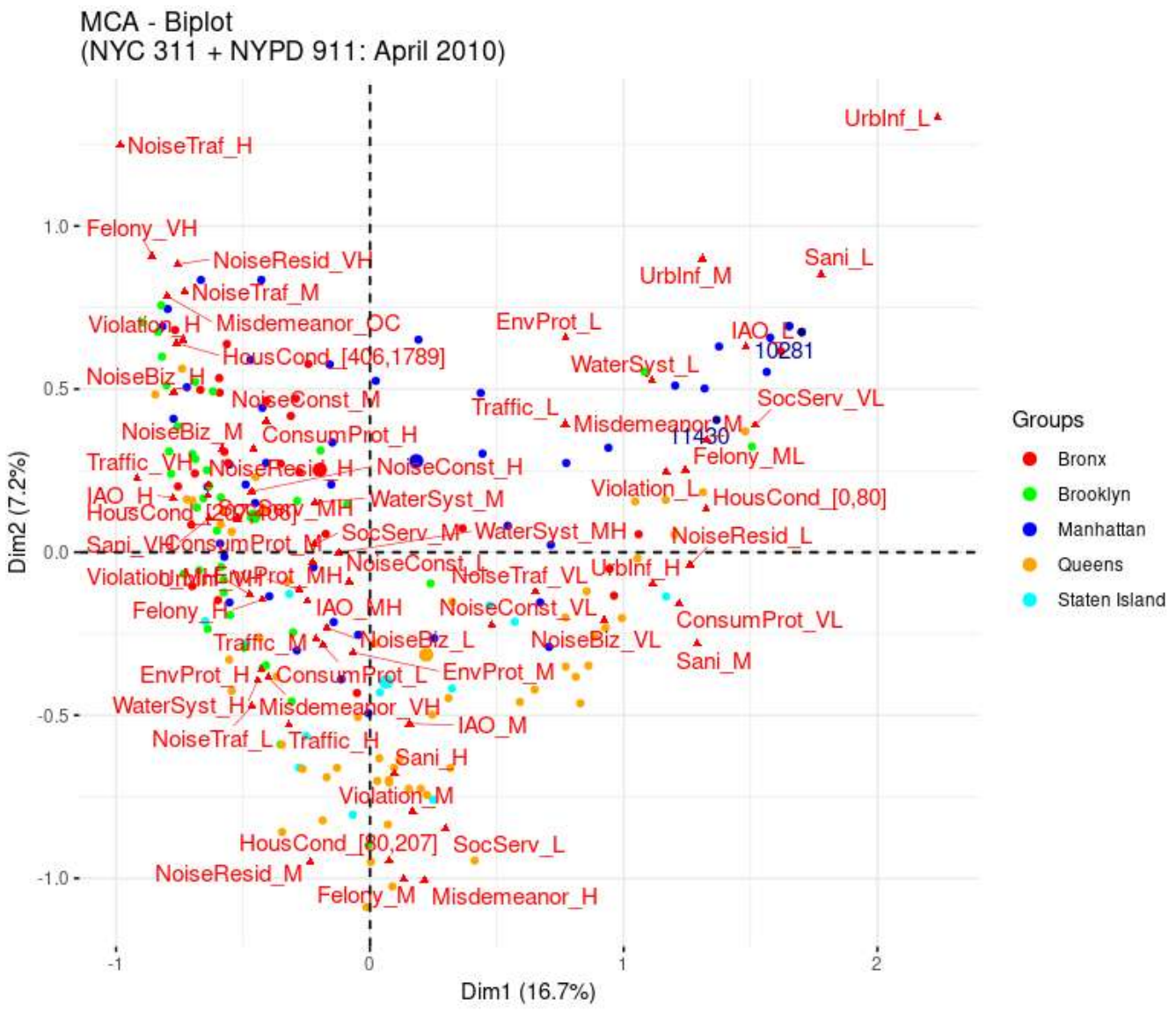
Crime factors in purple
Other factors in beige
ZIP code individuals are color coded according to borough.


a) Variables' projection in PC1-2 (all)
(MCA on April 2010 NYC data)
UrbInf_L
NoiseTraf_H
Felony_VH
NoiseResid_VH
NoiseTraf_M
Misdemeanor_OC
Violation_H
HousCond_[406,1789]
NoiseBiz_H
NoiseConst_M
NoiseBiz_M
ConsumProt_H
Traffic_VH
NoiseResid_H
NoiseConst_H
IAO_H
SocServ_MH
WaterSyst_M
HousCond_[207,406]
WaterSyst_MH
Sani_VH
EnvProt_MH
Violation_MH
Felony_H
UrbInf_VH
Traffic_M
EnvProt_H
WaterSyst_H
Misdemeanor_VH
NoiseTraf_L
Traffic_H
IAO_M
Sani_H
Violation_M
HousCond_[80,207]
SocServ_L
NoiseResid_M
Felony_M
Misdemeanor_H
UrbInf_M
Sani_L
EnvProt_L
IAO_L
WaterSyst_L
Traffic_L
SocServ_VL
Misdemeanor_M
Felony_ML
Violation_L
HousCond_[0,80]
NoiseResid_L
UrbInf_H
NoiseTraf_VL
NoiseConst_VL
ConsumProt_VL
NoiseBiz_VL
Sani_M
ConsumProt_L
EnvProt_M
Col.
Crime
SRCs
Dim2 (7.2%)
Dim1 (16.7%)
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
2

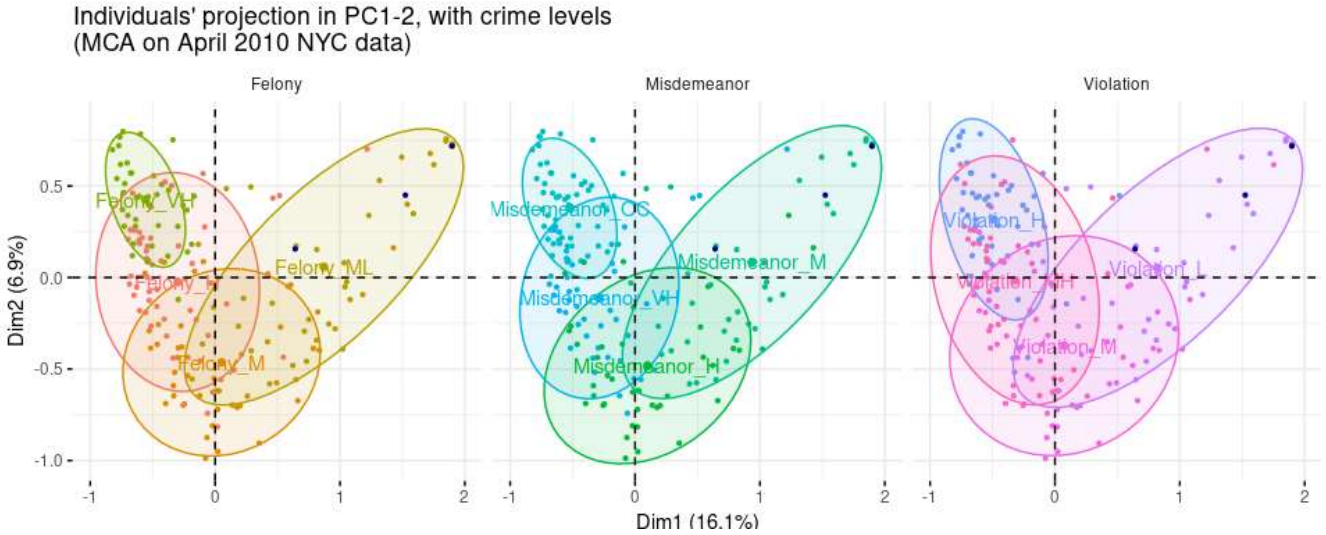

b) Variables' projection in PC1-2 (cos2>0.2)
(MCA on April 2010 NYC data)
UrbInf_L
Felony_VH
NoiseResid_VH
NoiseTraf_M
Misdemeanor_OC
Violation_H
HousCond_[406,1789]
IAO_H
SocServ_MH
Sani_VH
UrbInf_VH
WaterSyst_H
Violation_M
NoiseResid_M
HousCond_[80,207]
Felony_M
Misdemeanor_H
Sani_L
IAO_L
EnvProt_L
WaterSyst_L
Traffic_L
SocServ_VL
Misdemeanor_M
Felony_ML
Violation_L
HousCond_[0,80]
NoiseResid_L
NoiseTraf_VL
ConsumProt_VL
NoiseBiz_VL
Col.
Crime
SRCs
Dim2 (7.2%)
Dim1 (16.7%)
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
2


a) Individuals' projection in PC1-2
(MCA on April 2010 NYC data)
Dim2 (6.9%)
Dim1 (16.1%)
0.5
0.0
-0.5
-1.0
0
1
2
cos2
0.6
0.4
0.2


b) Individuals' projection in PC1-2, by NYC borough
(MCA on April 2010 NYC data)
Dim2 (6.9%)
Dim1 (16.1%)
0.5
0.0
-0.5
-1.0
0
1
2
Groups
Bronx
Brooklyn
Manhattan
Queens
Staten Island

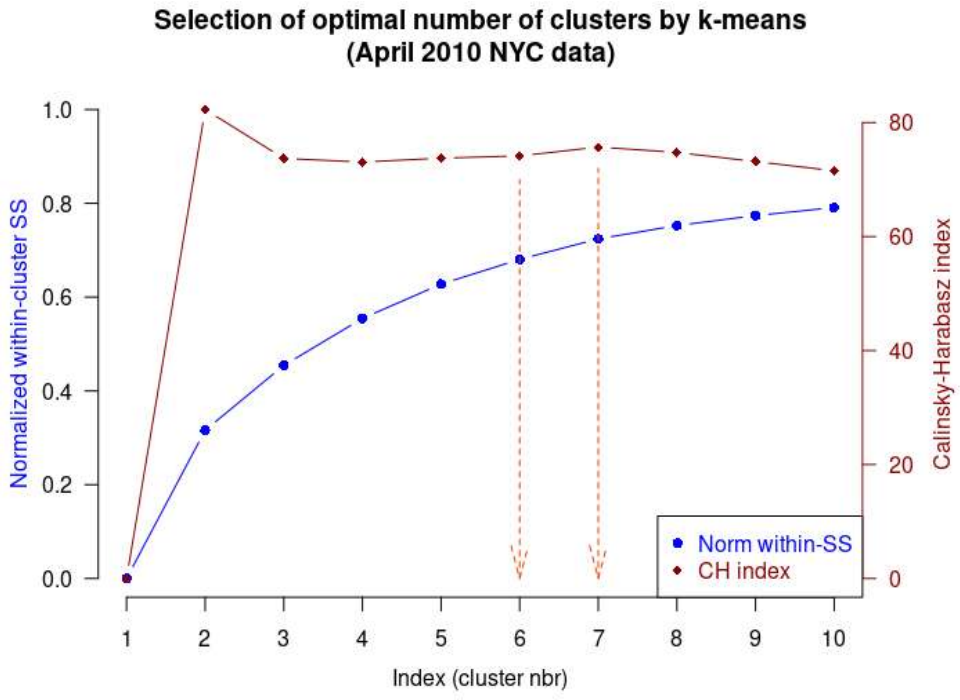

MCA - Biplot
(NYC 311 + NYPD 911: April 2010)
UrbInf_L
NoiseTraf_H
Felony_VH
NoiseResid_VH
NoiseTraf_M
Violation_H
Misdemeanor_OC
HousCond_[406,1789]
NoiseBiz_H
NoiseConst_M
NoiseBiz_M
ConsumProt_H
Traffic_VH
NoiseResid_H
NoiseConst_H
IAO_H
WaterSyst_M
HousCond_[207,406]
SocServ_M
WaterSyst_MH
Sani_VH
ConsumProt_M
NoiseConst_L
NoiseTraf_VL
UrbInf_H
Violation_MH
EnvProt_MH
Felony_H
IAO_MH
NoiseConst_VL
ConsumProt_VL
Traffic_M
NoiseBiz_L
NoiseBiz_VL
Sani_M
EnvProt_H
ConsumProt_L
EnvProt_M
WaterSyst_H
Misdemeanor_VH
IAO_M
NoiseTraf_L
Traffic_H
Sani_H
Violation_M
HousCond_[80,207]
SocServ_L
NoiseResid_M
Felony_M
Misdemeanor_H
EnvProt_L
UrbInf_M
Sani_L
IAO_L
10281
WaterSyst_L
SocServ_VL
Traffic_L
Misdemeanor_M
11430
Felony_ML
Violation_L
HousCond_[0,80]
NoiseResid_L
1.0
0.5
0.0
-0.5
-1.0
-1
0
1
2
Dim2 (7.2%)
Dim1 (16.7%)
Groups
Bronx
Brooklyn
Manhattan
Queens
Staten Island

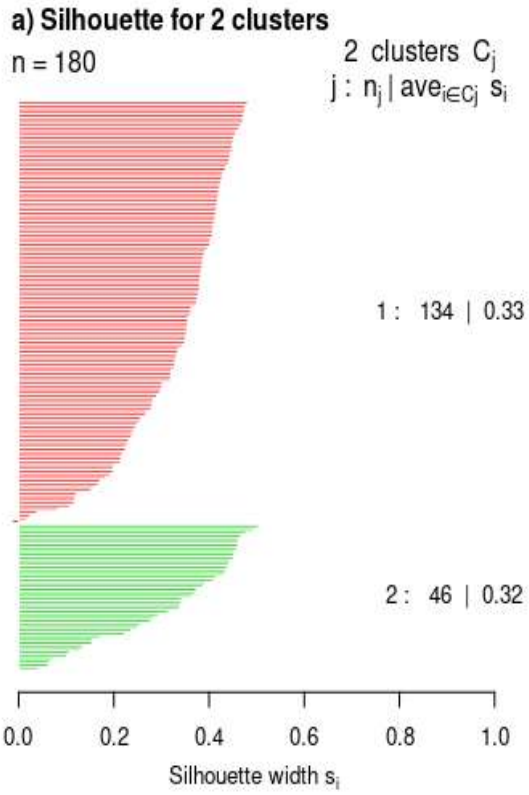
Identify ZIP codes in 2nd quadrant of PC12 var projection from MCA:



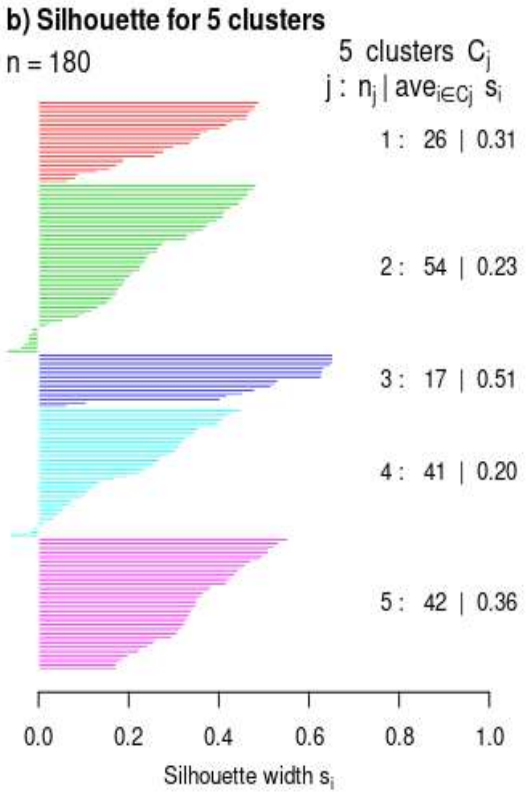

```
sum of all violation counts in 2nd quadrant: 2759
sum of all other violation counts: 2292

sum of all misdemeanor counts in 2nd quadrant: 15214
sum of all other misdemeanor counts: 10399

sum of all felony counts in 2nd quadrant: 6390
sum of all other felony counts: 11542
```

Clustering analysis:

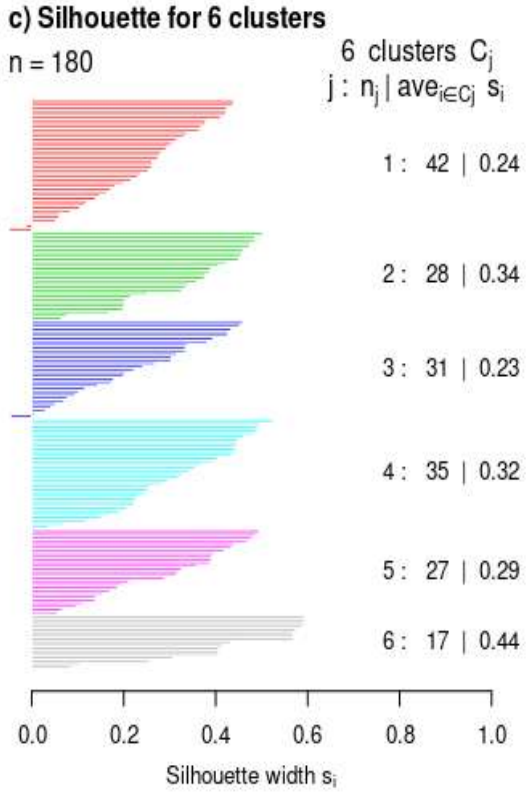




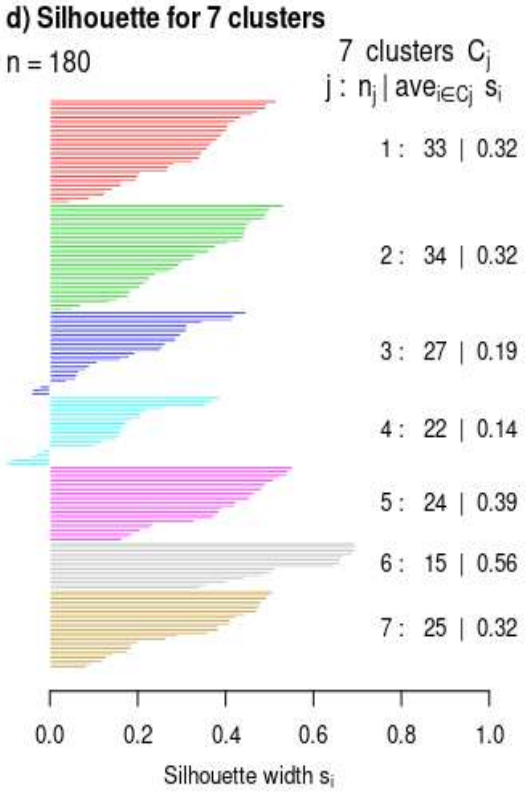
Average silhouette width : 0.33



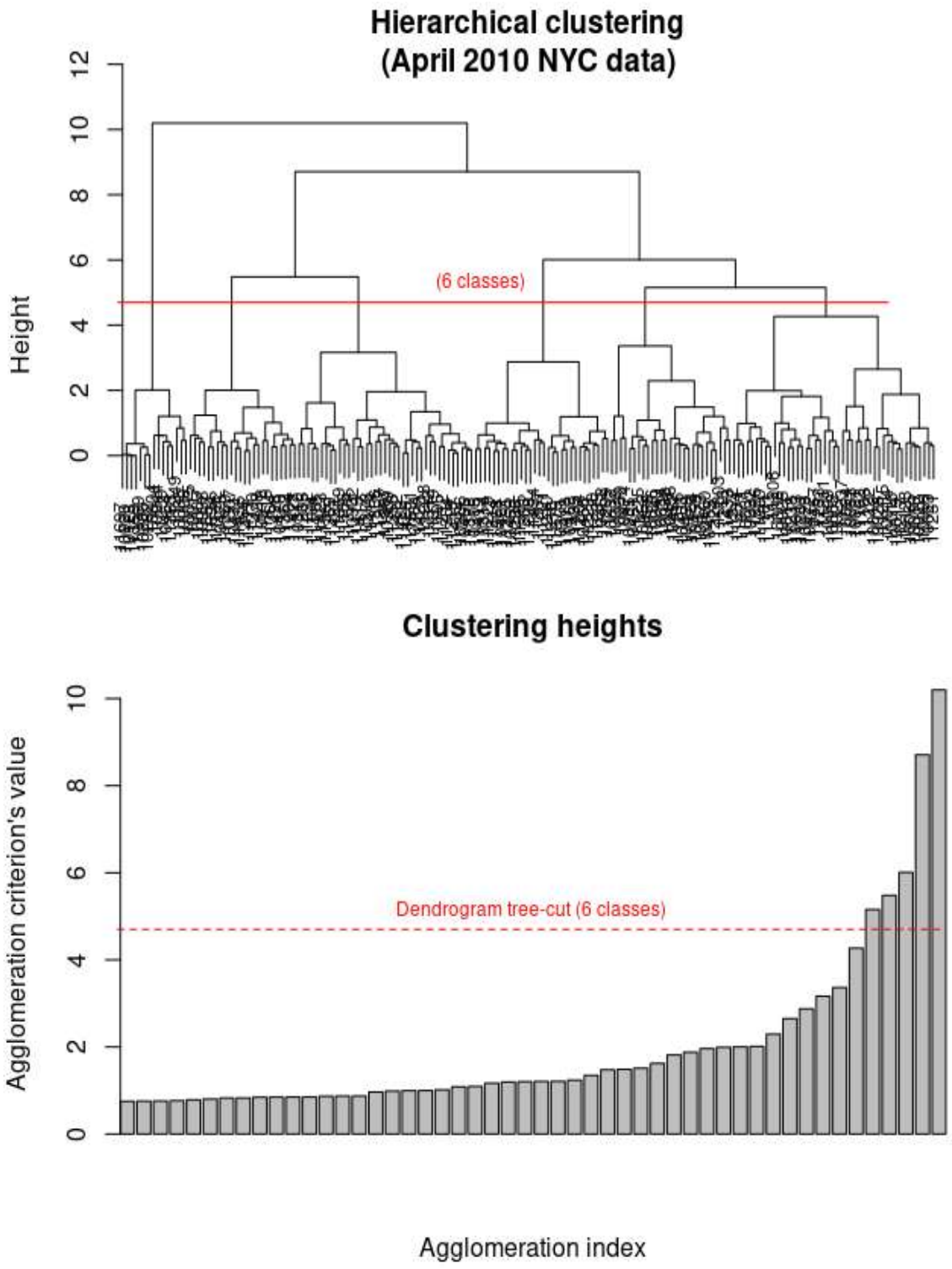
Average silhouette width : 0.29



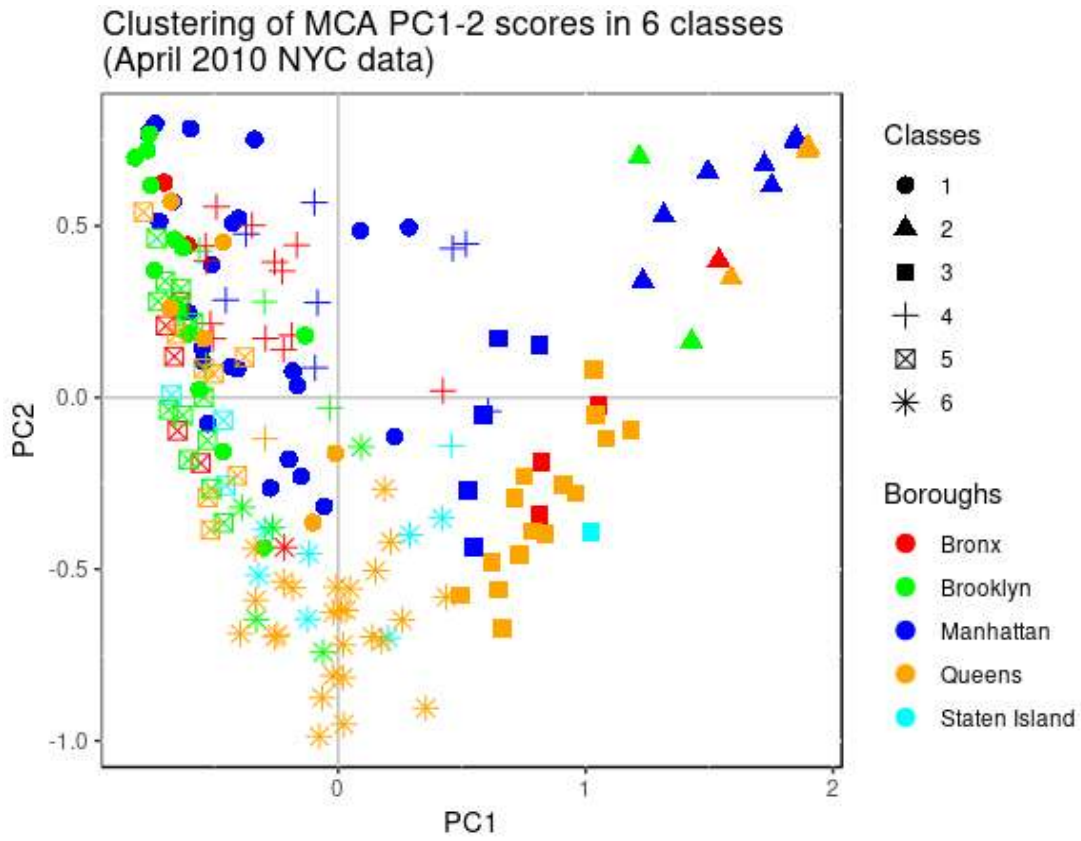
Average silhouette width : 0.3

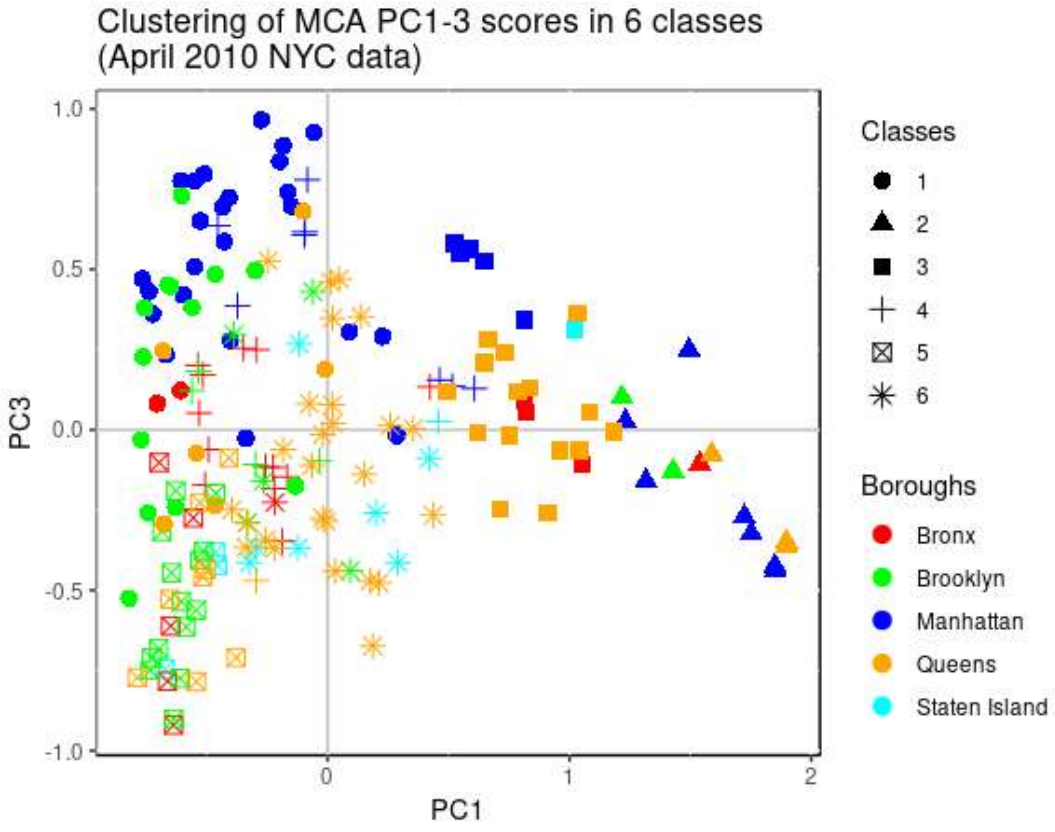


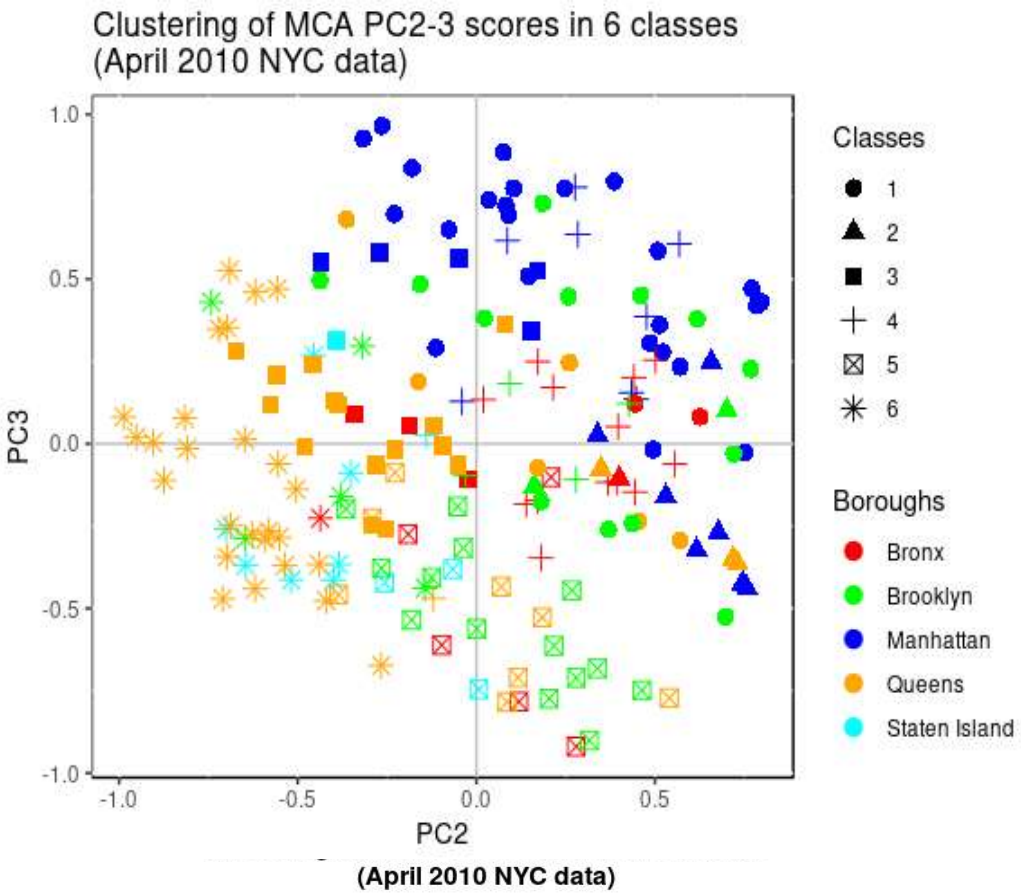
Average silhouette width : 0.31



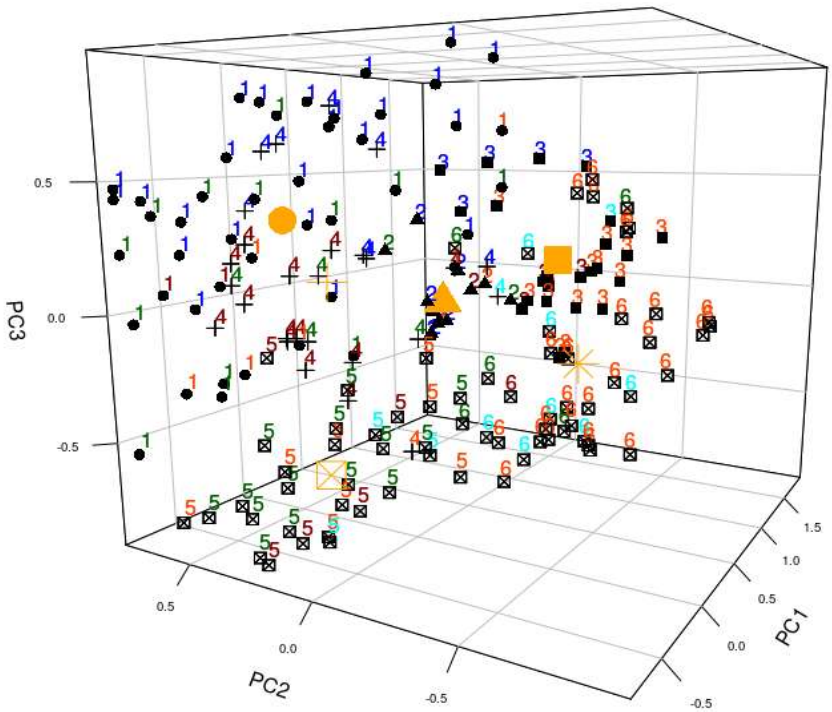
Clustering quality index for 6 classes, I_b=65.98


Clustering of MCA PC1-2 scores in 6 classes
(April 2010 NYC data)
PC2
PC1
0.5
0.0
-0.5
-1.0
0
1
2
Classes
1
2
3
4
5
6
Boroughs
Bronx
Brooklyn
Manhattan
Queens
Staten Island


Clustering of MCA PC1-3 scores in 6 classes
(April 2010 NYC data)
PC3
PC1
1.0
0.5
0.0
-0.5
-1.0
0
1
2
Classes
1
2
3
4
5
6
Boroughs
Bronx
Brooklyn
Manhattan
Queens
Staten Island



(April 2010 NYC data)



(phi=10º, theta=-60º)

Contributions of each borough to inertia over 'nd' = 5 dimensions:

```
Borough: Bronx
Boroughs' % (normalized) contribution to inertia:  12.9 %

Borough: Brooklyn
Boroughs' % (normalized) contribution to inertia:  20.2 %

Borough: Manhattan
Boroughs' % (normalized) contribution to inertia:  33.9 %

Borough: Queens
Boroughs' % (normalized) contribution to inertia:  28.9 %

Borough: Staten Island
Boroughs' % (normalized) contribution to inertia:  4.2 %
```

For more detail:

```
Borough:  Bronx
 Number of ZIP codes:  25
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 4.9
      5_dim
10451   0.1
10452   0.1
10453   0.3
10454   0.1
10455   0.2
10456   0.2
10457   0.3
10458   0.3
10459   0.1
10460   0.2
10461   0.2
10462   0.2
10463   0.1
10464   0.5
10465   0.1
10466   0.3
10467   0.3
10468   0.1
10469   0.2
10470   0.3
10471   0.2
10472   0.2
10473   0.1
10474   0.1
10475   0.1

Borough:  Brooklyn
 Number of ZIP codes:  38
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 7.7
      5_dim
11201   0.2
11203   0.2
11204   0.2
11205   0.1
11206   0.3
```

	5_dim
11207	0.1
11208	0.2
11209	0.2
11210	0.1
11211	0.4
11212	0.1
11213	0.1
11214	0.1
11215	0.3
11216	0.2
11217	0.2
11218	0.1
11219	0.1
11220	0.2
11221	0.3
11222	0.2
11223	0.2
11224	0.2
11225	0.1
11226	0.3
11228	0.1
11229	0.2
11230	0.2
11231	0.1
11232	0.3
11233	0.2
11234	0.2
11235	0.2
11236	0.2
11237	0.1
11238	0.2
11239	0.4
11249	0.6

Borough: Manhattan
Number of ZIP codes: 46
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 12.9

	5_dim
10001	0.1
10002	0.4
10003	0.2
10004	0.5
10005	0.5
10006	0.3
10007	0.1
10009	0.3
10010	0.2
10011	0.2
10012	0.2
10013	0.2
10014	0.3
10016	0.2
10017	0.1
10018	0.2
10019	0.3
10020	0.8
10021	0.0
10022	0.2
10023	0.2
10024	0.1

```
10025   0.1
10026   0.2
10027   0.3
10028   0.3
10029   0.2
10030   0.2
10031   0.1
10032   0.1
10033   0.1
10034   0.1
10035   0.1
10036   0.1
10037   0.4
10038   0.1
10039   0.4
10040   0.2
10044   0.7
10065   0.1
10069   0.7
10075   0.2
10123   0.8
10128   0.2
10280   0.8
10282   0.8

Borough:  Queens
 Number of ZIP codes:  59
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 11
      5_dim
11004   0.4
11101   0.1
11102   0.0
11103   0.2
11104   0.2
11105   0.1
11106   0.2
11354   0.1
11355   0.2
11356   0.2
11357   0.1
11358   0.1
11359   0.8
11360   0.2
11361   0.1
11362   0.2
11363   0.5
11364   0.2
11365   0.1
11366   0.2
11367   0.2
11368   0.1
11369   0.1
11370   0.2
11372   0.1
11373   0.2
11374   0.1
11375   0.1
11377   0.1
11378   0.1
11379   0.1
```

```
11385   0.3
11411   0.2
11412   0.2
11413   0.1
11414   0.2
11415   0.3
11416   0.2
11417   0.1
11418   0.2
11419   0.1
11420   0.1
11421   0.2
11422   0.1
11423   0.2
11426   0.3
11427   0.3
11428   0.2
11429   0.1
11432   0.1
11433   0.1
11434   0.1
11435   0.1
11436   0.2
11691   0.1
11692   0.2
11693   0.2
11694   0.1
11697   0.8

Borough:  Staten Island
 Number of ZIP codes:  12
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 1.6
      5_dim
10301   0.1
10302   0.1
10303   0.1
10304   0.1
10305   0.1
10306   0.1
10307   0.3
10308   0.1
10309   0.1
10310   0.1
10312   0.2
10314   0.2
```

Contributions of each class to inertia over 'nd' = 5 dimensions:

```
Cluster's 1 normalized contribution to inertia:  22.8 %
Cluster's 2 normalized contribution to inertia:  24.9 %
Cluster's 3 normalized contribution to inertia:  13.1 %
Cluster's 4 normalized contribution to inertia:  12.3 %
Cluster's 5 normalized contribution to inertia:  13.4 %
Cluster's 6 normalized contribution to inertia:  13.4 %
```

For more detail:

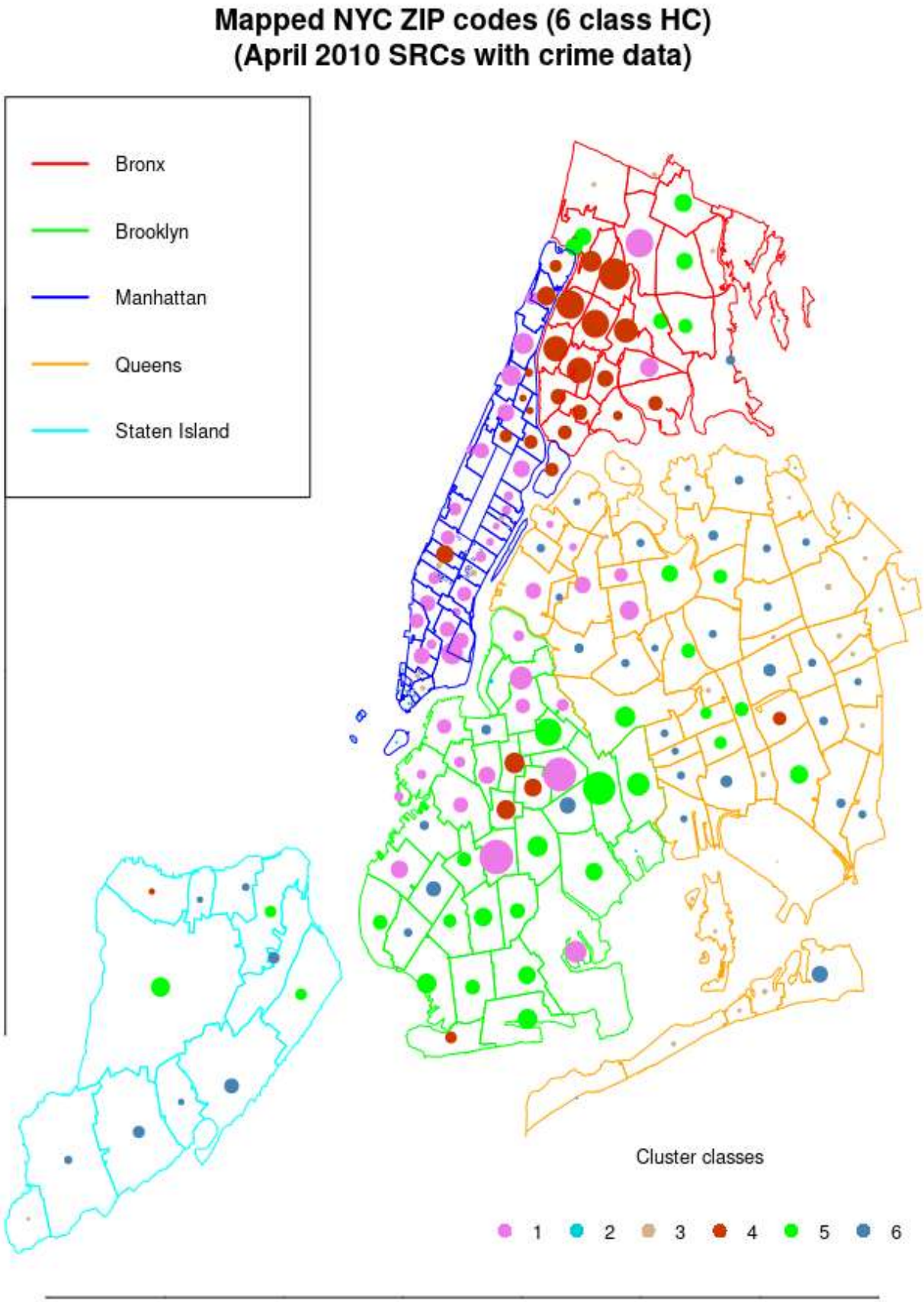
```
Cluster:  1
```

```
 Number of ZIP codes:  45
Clusters' (non-normalized) % contribution to inertia: 8.7
      IEP_over_5_dim
10001            0.1
10002            0.4
10003            0.2
10009            0.3
10010            0.2
10011            0.2
10012            0.2
10013            0.2
10014            0.3
10016            0.2
10019            0.3
10021            0.0
10022            0.2
10023            0.2
10024            0.1
10025            0.1
10027            0.3
10028            0.3
10029            0.2
10031            0.1
10032            0.1
10033            0.1
10065            0.1
10128            0.2
10467            0.3
10472            0.2
11101            0.1
11102            0.0
11103            0.2
11201            0.2
11206            0.3
11211            0.4
11215            0.3
11217            0.2
11220            0.2
11222            0.2
11226            0.3
11231            0.1
11233            0.2
11234            0.2
11237            0.1
11238            0.2
11372            0.1
11373            0.2
11377            0.1

Cluster:  2
 Number of ZIP codes:  15
Clusters' (non-normalized) % contribution to inertia: 9.5
      IEP_over_5_dim
10004            0.5
10005            0.5
10006            0.3
```

```
10020            0.8
10044            0.7
10069            0.7
10123            0.8
10280            0.8
10282            0.8
10464            0.5
11239            0.4
11249            0.6
11359            0.8
11363            0.5
11697            0.8

Cluster:  3
 Number of ZIP codes:  24
Clusters' (non-normalized) % contribution to inertia: 5
      IEP_over_5_dim
10007            0.1
10017            0.1
10018            0.2
10038            0.1
10075            0.2
10307            0.3
10470            0.3
10471            0.2
10475            0.1
11004            0.4
11360            0.2
11362            0.2
11364            0.2
11366            0.2
11370            0.2
11411            0.2
11415            0.3
11426            0.3
11427            0.3
11428            0.2
11436            0.2
11692            0.2
11693            0.2
11694            0.1

Cluster:  4
 Number of ZIP codes:  27
Clusters' (non-normalized) % contribution to inertia: 4.7
      IEP_over_5_dim
10026            0.2
10030            0.2
10034            0.1
10035            0.1
10036            0.1
10037            0.4
10039            0.4
10040            0.2
10303            0.1
10451            0.1
```

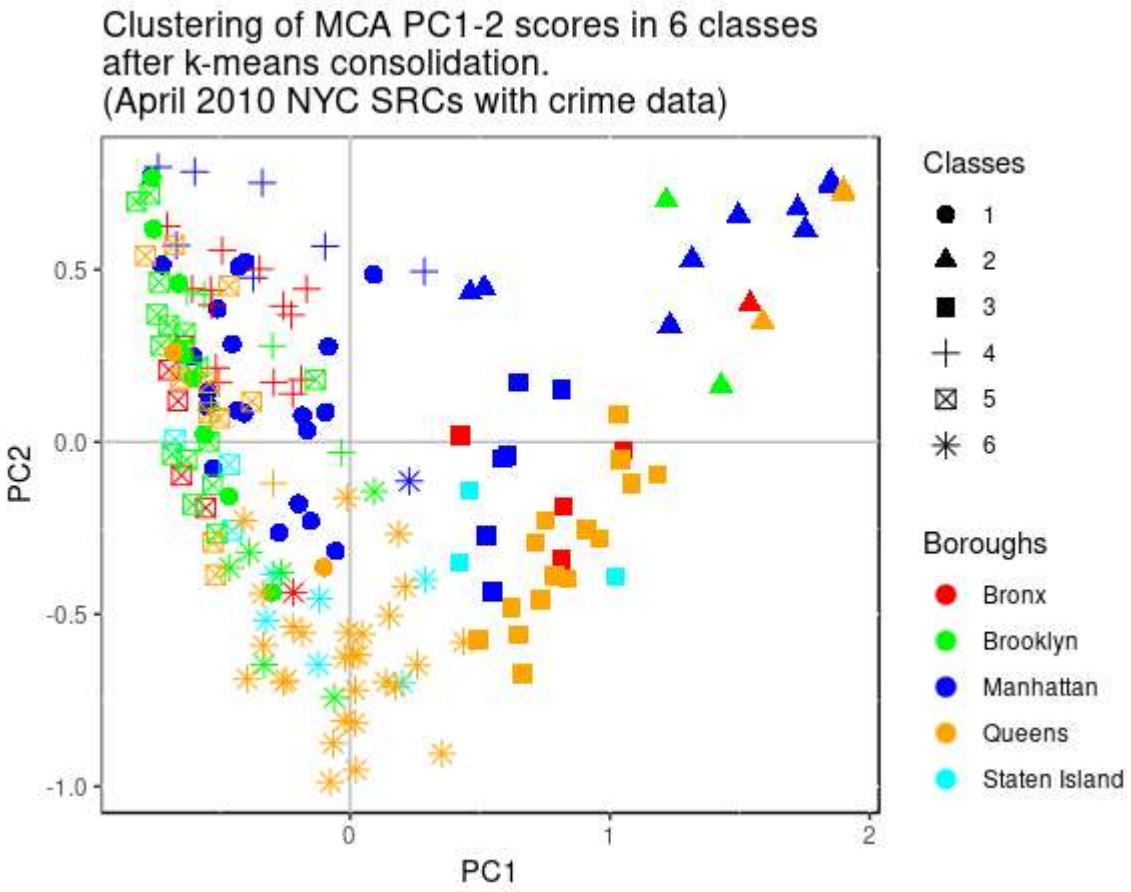
```
10452            0.1
10453            0.3
10454            0.1
10455            0.2
10456            0.2
10457            0.3
10458            0.3
10459            0.1
10460            0.2
10468            0.1
10473            0.1
10474            0.1
11213            0.1
11216            0.2
11224            0.2
11225            0.1
11433            0.1

Cluster:  5
 Number of ZIP codes:  30
Clusters' (non-normalized) % contribution to inertia: 5.1
      IEP_over_5_dim
10301            0.1
10305            0.1
10314            0.2
10461            0.2
10462            0.2
10463            0.1
10466            0.3
10469            0.2
11203            0.2
11204            0.2
11207            0.1
11208            0.2
11209            0.2
11210            0.1
11214            0.1
11218            0.1
11221            0.3
11223            0.2
11229            0.2
11230            0.2
11235            0.2
11236            0.2
11355            0.2
11368            0.1
11375            0.1
11385            0.3
11418            0.2
11419            0.1
11434            0.1
11435            0.1

Cluster:  6
 Number of ZIP codes:  39
Clusters' (non-normalized) % contribution to inertia: 5.1
```

	IEP_over_5_dim
10302	0.1
10304	0.1
10306	0.1
10308	0.1
10309	0.1
10310	0.1
10312	0.2
10465	0.1
11104	0.2
11105	0.1
11106	0.2
11205	0.1
11212	0.1
11219	0.1
11228	0.1
11232	0.3
11354	0.1
11356	0.2
11357	0.1
11358	0.1
11361	0.1
11365	0.1
11367	0.2
11369	0.1
11374	0.1
11378	0.1
11379	0.1
11412	0.2
11413	0.1
11414	0.2
11416	0.2
11417	0.1
11420	0.1
11421	0.2
11422	0.1
11423	0.2
11429	0.1
11432	0.1
11691	0.1

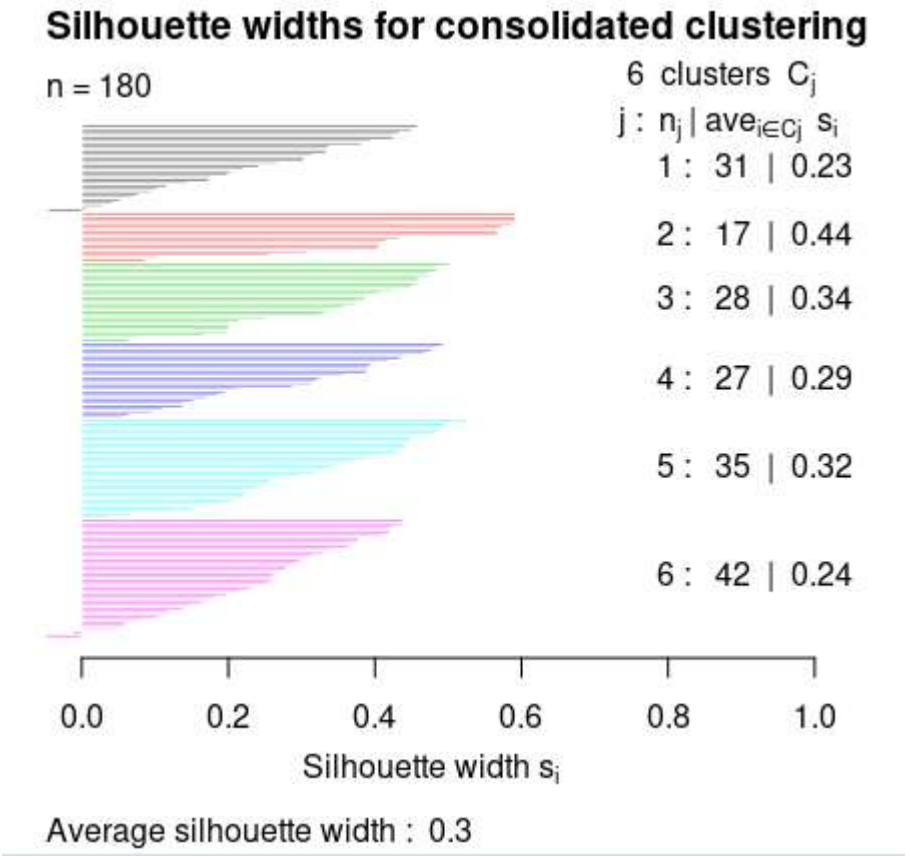

Mapped NYC ZIP codes (6 class HC)
(April 2010 SRCs with crime data)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Cluster classes
1
2
3
4
5
6

HC clustering with k-means consolidation:

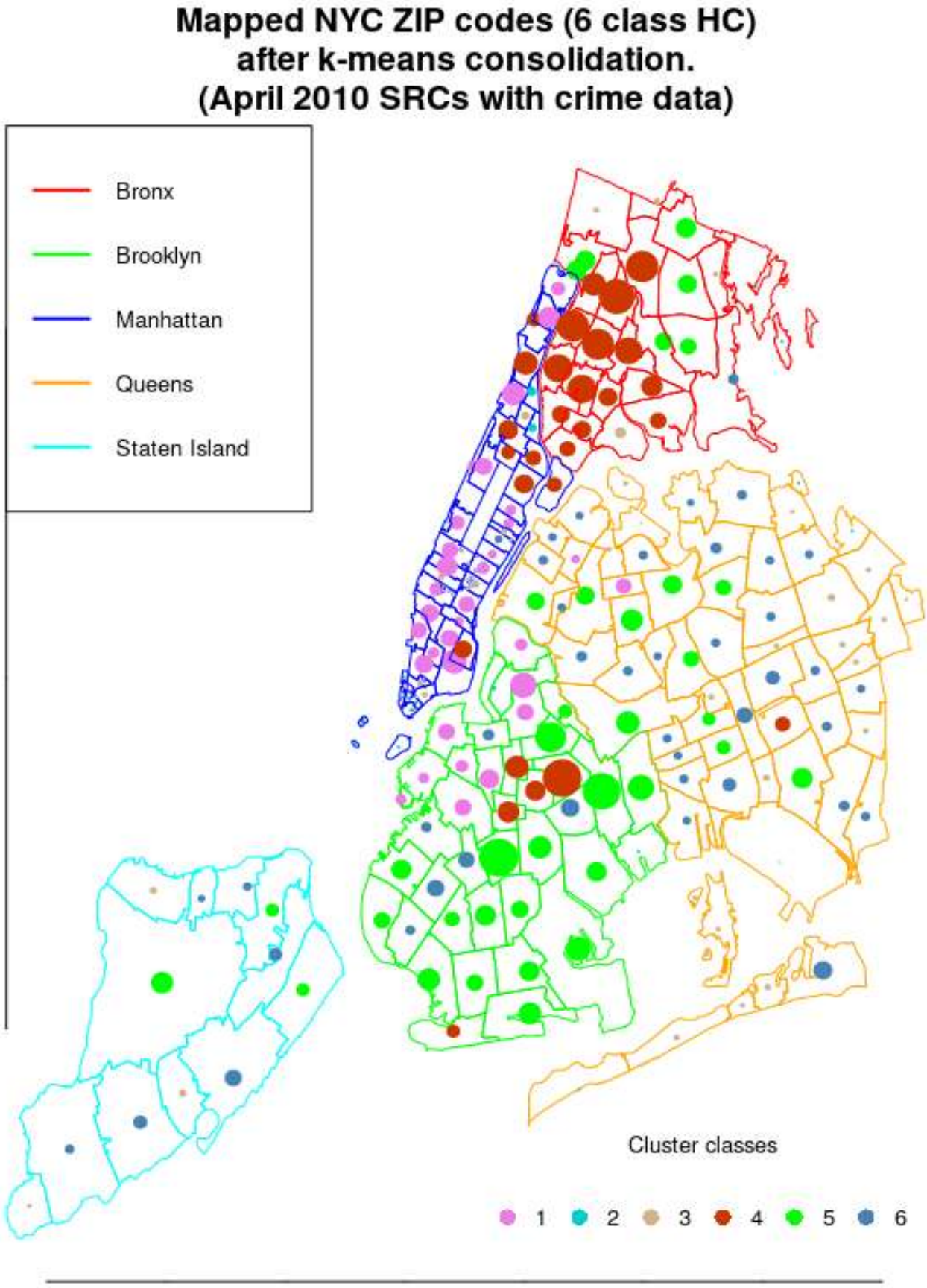
Clustering quality Index for 6 classes, after k-means consolidation, Ib = 68.15



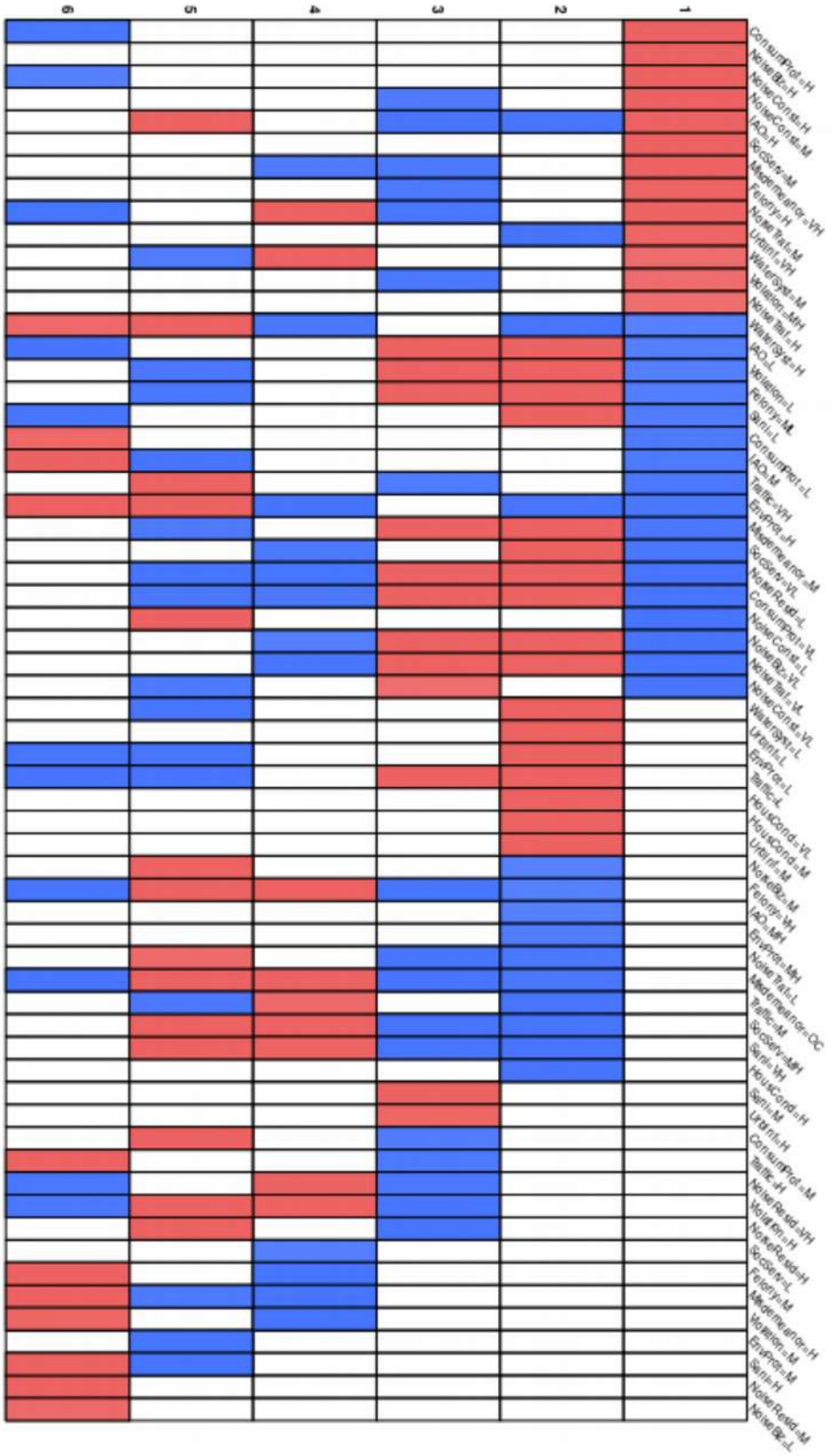
Cluster's 1 normalized contribution to inertia: 16.3 %
Cluster's 2 normalized contribution to inertia: 27 %
Cluster's 3 normalized contribution to inertia: 14.4 %
Cluster's 4 normalized contribution to inertia: 12.6 %
Cluster's 5 normalized contribution to inertia: 13.6 %
Cluster's 6 normalized contribution to inertia: 16 %

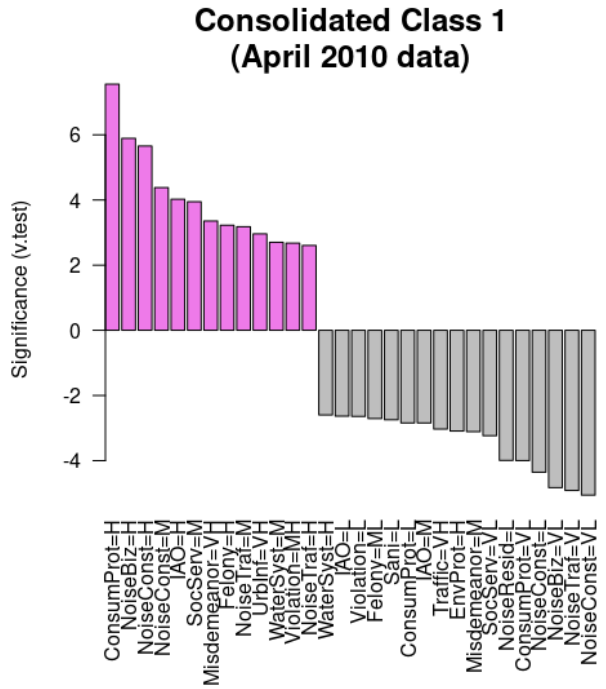


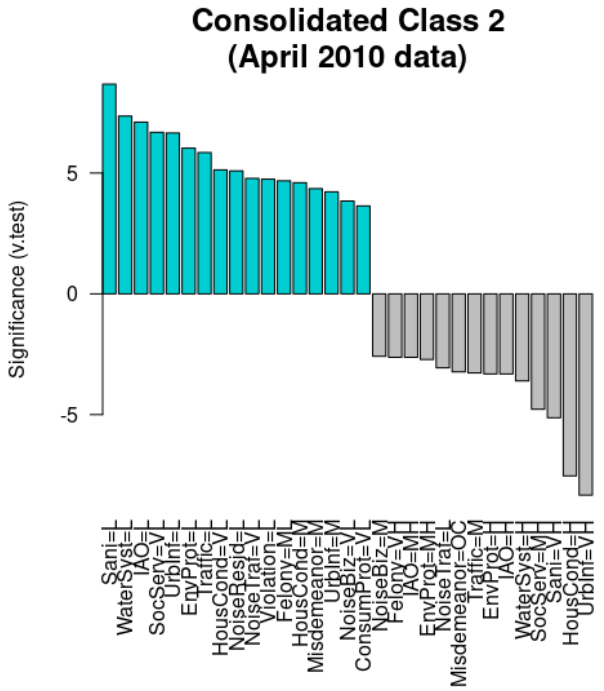

```
Class  1 's most significant modalities:
 ConsumProt=H NoiseBiz=H NoiseConst=H NoiseConst=M IAO=H SocServ=M Misdemeanor=VH
Felony=H NoiseTraf=M UrbInf=VH WaterSyst=M Violation=MH NoiseTraf=H

Class  2 's most significant modalities:
 Sani=L WaterSyst=L IAO=L SocServ=VL UrbInf=L EnvProt=L Traffic=L HousCond=VL
NoiseResid=L NoiseTraf=VL Violation=L Felony=ML HousCond=M Misdemeanor=M UrbInf=M
NoiseBiz=VL ConsumProt=VL

Class  3 's most significant modalities:
 NoiseResid=L Sani=M NoiseTraf=VL Misdemeanor=M ConsumProt=VL NoiseBiz=VL
Felony=ML Traffic=L Violation=L IAO=L UrbInf=H NoiseConst=VL

Class  4 's most significant modalities:
 Misdemeanor=OC Felony=VH Violation=H SocServ=MH NoiseResid=VH NoiseTraf=M
WaterSyst=M Sani=VH Traffic=M

Class  5 's most significant modalities:
 Traffic=VH EnvProt=H Sani=VH WaterSyst=H IAO=H NoiseResid=H SocServ=MH Felony=VH
ConsumProt=M Misdemeanor=OC NoiseConst=L Violation=H NoiseBiz=M NoiseTraf=L

Class  6 's most significant modalities:
 Misdemeanor=H NoiseResid=M Traffic=H Felony=M IAO=M Sani=H EnvProt=H Violation=M
NoiseBiz=L ConsumProt=L WaterSyst=H
```

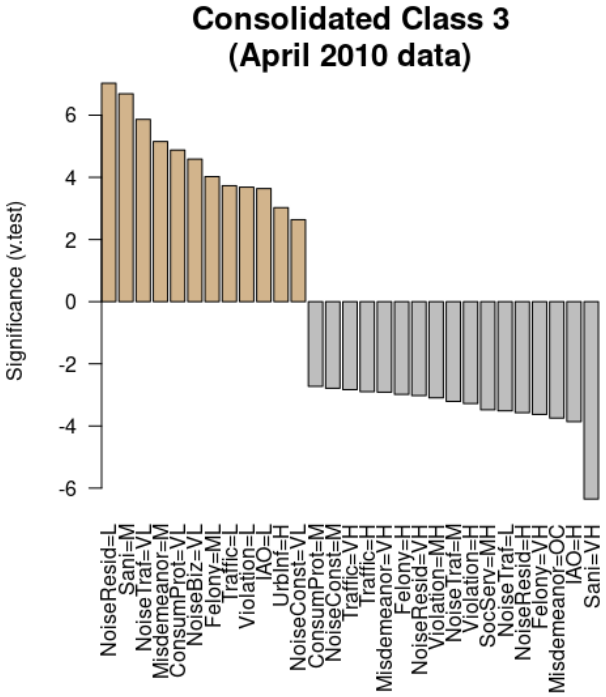

Mapped NYC ZIP codes (6 class HC)
after k-means consolidation.
(April 2010 SRCs with crime data)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
Cluster classes
1
2
3
4
5
6

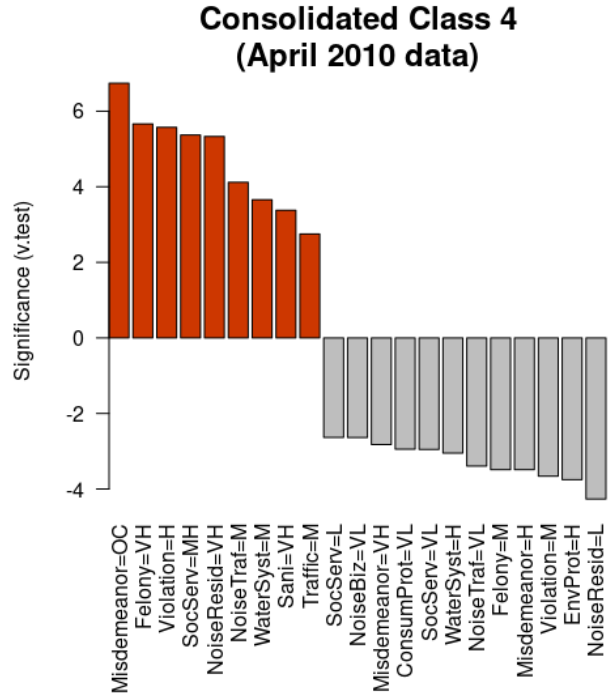
Clustering interpretation:

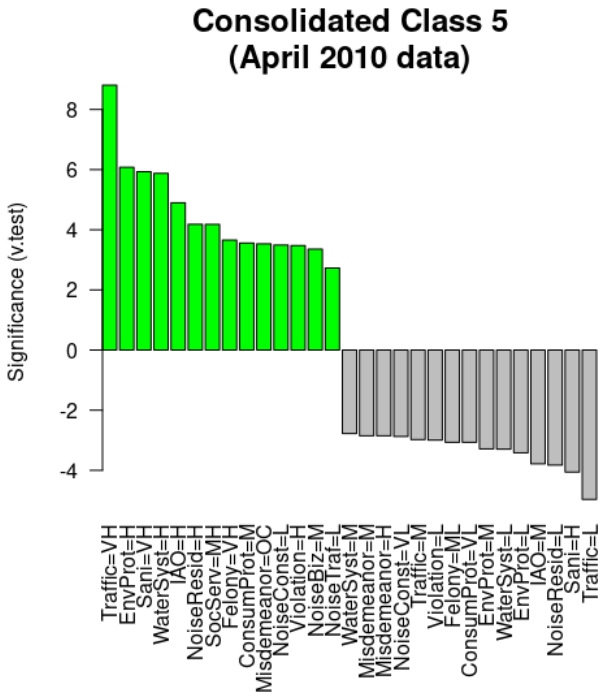


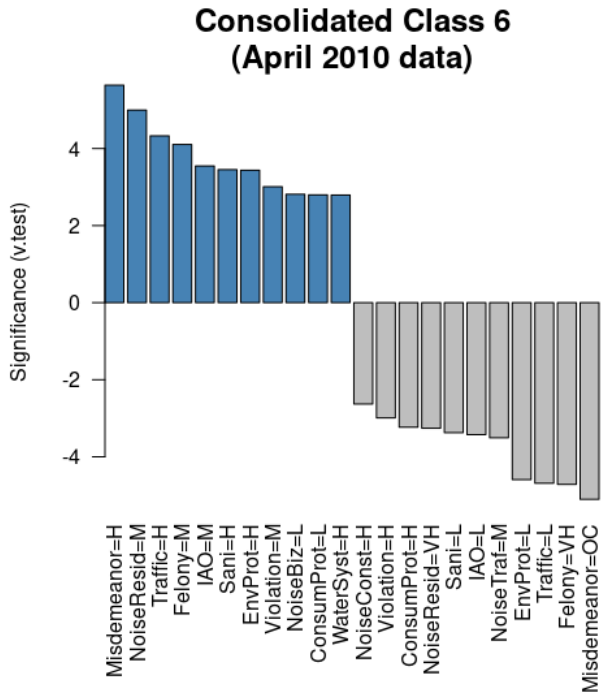

Consolidated Class 1
(April 2010 data)
Significance (v.test)
6
4
2
0
-2
-4
ConsumProt=H
NoiseBiz=H
NoiseConst=H
NoiseConst=M
IAO=H
SocServ=M
Misdemeanor=VH
Felony=H
NoiseTraf=M
UrbInf=VH
WaterSyst=M
Violation=MH
NoiseTraf=H
WaterSyst=H
IAO=L
Violation=L
Felony=ML
Sani=L
ConsumProt=L
IAO=M
Traffic=VH
EnvProt=H
Misdemeanor=M
SocServ=VL
NoiseResid=L
ConsumProt=VL
NoiseConst=L
NoiseBiz=VL
NoiseTraf=VL
NoiseConst=VL


Consolidated Class 2
(April 2010 data)
Significance (v.test)
5
0
-5
Sani=L
WaterSyst=L
IAO=L
SocServ=VL
UrbInf=L
EnvProt=L
Traffic=L
HousCond=VL
NoiseResid=L
NoiseTraf=VL
Violation=L
Felony=ML
HousCond=M
Misdemeanor=M
UrbInf=M
NoiseBiz=VL
ConsumProt=VL
NoiseBiz=M
Felony=VH
IAO=MH
EnvProt=MH
NoiseTraf=L
Misdemeanor=OC
Traffic=M
EnvProt=H
IAO=H
WaterSyst=H
SocServ=MH
Sani=VH
HousCond=H
UrbInf=VH

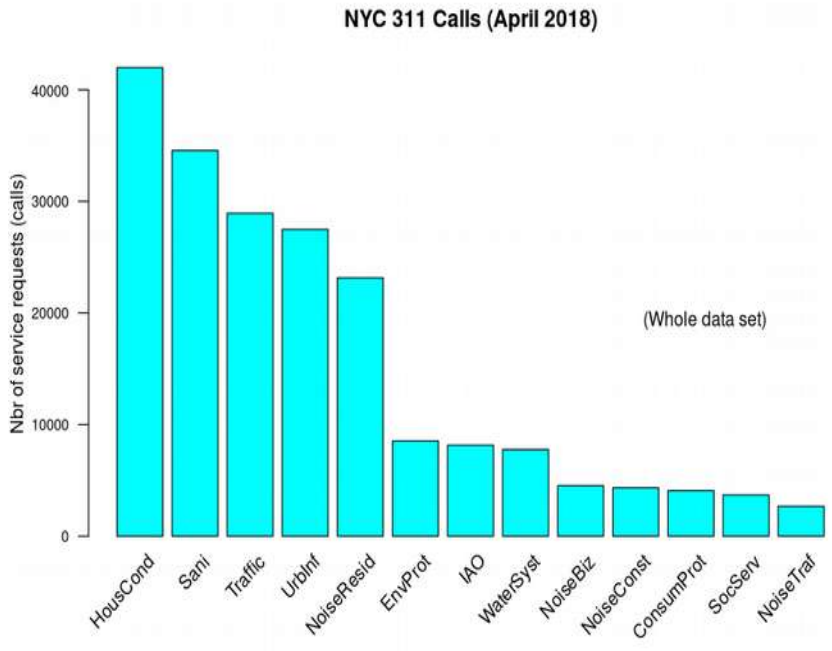

Consolidated Class 3
(April 2010 data)
Significance (v.test)
6
4
2
0
-2
-4
-6
NoiseResid=L
Sani=M
NoiseTraf=VL
Misdemeanor=M
ConsumProt=VL
NoiseBiz=VL
Felony=ML
Traffic=L
Violation=L
IAO=L
UrbInf=H
NoiseConst=VL
ConsumProt=M
NoiseConst=M
Traffic=VH
Traffic=H
Misdemeanor=VH
Felony=H
NoiseResid=VH
Violation=MH
NoiseTraf=M
Violation=H
SocServ=MH
NoiseTraf=L
NoiseResid=H
Felony=VH
Misdemeanor=OC
IAO=H
Sani=VH


Consolidated Class 4
(April 2010 data)
Significance (v.test)
6
4
2
0
-2
-4
Misdemeanor=OC
Felony=VH
Violation=H
SocServ=MH
NoiseResid=VH
NoiseTraf=M
WaterSyst=M
Sani=VH
Traffic=M
SocServ=L
NoiseBiz=VL
Misdemeanor=VH
ConsumProt=VL
SocServ=VL
WaterSyst=H
NoiseTraf=VL
Felony=M
Misdemeanor=H
Violation=M
EnvProt=H
NoiseResid=L


Consolidated Class 5
(April 2010 data)
Significance (v.test)
8
6
4
2
0
-2
-4
Traffic=VH
EnvProt=H
Sani=VH
WaterSyst=H
IAO=H
NoiseResid=H
SocServ=MH
Felony=VH
ConsumProt=M
Misdemeanor=OC
NoiseConst=L
Violation=H
NoiseBiz=M
NoiseTraf=L
WaterSyst=M
Misdemeanor=M
Misdemeanor=H
NoiseConst=VL
Traffic=M
Violation=L
Felony=ML
ConsumProt=VL
EnvProt=M
WaterSyst=L
EnvProt=L
IAO=M
NoiseResid=L
Sani=H
Traffic=L

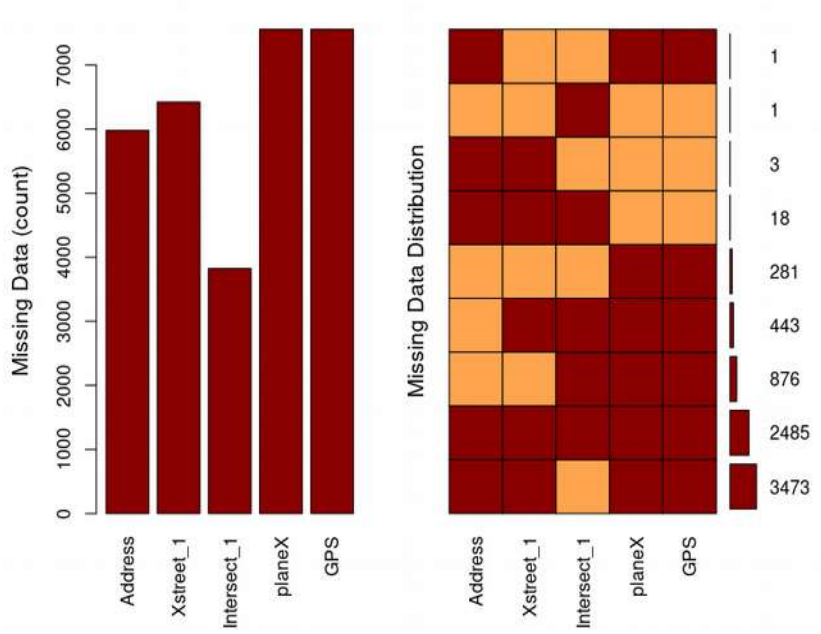

Consolidated Class 6
(April 2010 data)
Significance (v.test)
4
2
0
-2
-4
Misdemeanor=H
NoiseResid=M
Traffic=H
Felony=M
IAO=M
Sani=H
EnvProt=H
Violation=M
NoiseBiz=L
ConsumProt=L
WaterSyst=H
NoiseConst=H
Violation=H
ConsumProt=H
NoiseResid=VH
Sani=L
IAO=L
NoiseTraf=M
EnvProt=L
Traffic=L
Felony=VH
Misdemeanor=OC

Appendix G: Analytical results summary for the period April 2018



obs w/ missing ZIPs: 7,581
obs w/ missing GPS coords: 16,632
obs w/ missing ZIP and GPS coords: 7,759
obs w/ missing ZIP, Address, GPS coords 5,959

For all obs with missing ZIP
- 2485 obs miss all geoloc info



Missings per variable:

Variable	Count
Address	5980
Xstreet_1	6422
Intersect_1	3823
planeX	7559
GPS	7559

Missings in combinations of variables:

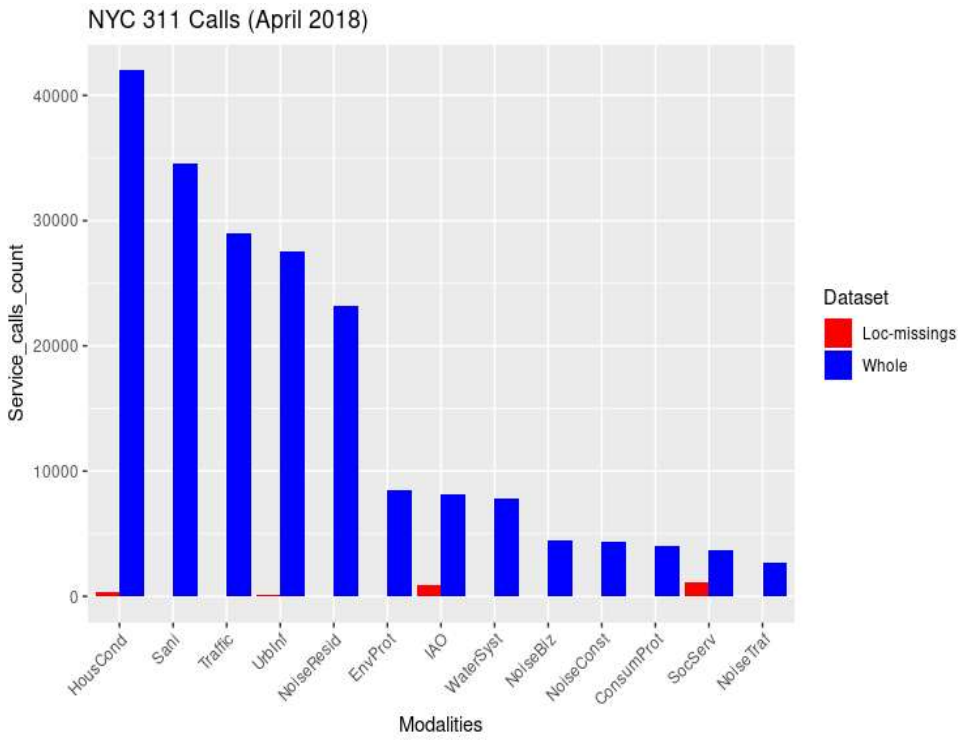
Combinations	Count	Percent
0:0:0:1:1	281	3.70663501
0:0:1:0:0	1	0.01319087
0:0:1:1:1	876	11.55520380
0:1:1:1:1	443	5.84355626
1:0:0:1:1	1	0.01319087
1:1:0:0:0	3	0.03957262
1:1:0:1:1	3473	45.81189817
1:1:1:0:0	18	0.23743569
1:1:1:1:1	2485	32.77931671

mod_allmissing

SocServ	IAO	HousCond	UrbInf	Sani	ConsumProt	EnvProt	NoiseBiz	NoiseConst	NoiseResid
1122	903	313	107	38	1	1	0	0	0

NoiseTraf	Traffic	WaterSyst
0	0	0

2485 (32.8% of all obs missing a ZIP code) have no geolocation information. Compare the modality distribution of those complaints, frequency wise, with those of the whole data set.



Fraction of modality 'SocServ': 30.4 %
Fraction of modality 'IAO': 11.1 %
Fraction of modality 'HousCond': 0.8 %

The Chi square test of independence of ZIP code observations and SRCs leads to a clear rejection of H_0.

5181 observations are fed to the Google API in an attempt to attribute a ZIP code and GPS coordinates to their corresponding event. 5354 Google Cloud geocode API queries are made as a result, at a cost of 26.77€). Of the 5181 observations 347 are left without GPS and ZIP code.

```
> table(protoY[imputeTo99999_idx,3])  # before "99999" imputation
ConsumProt    EnvProt   HousCond  IAO   NoiseBiz NoiseConst NoiseResid  NoiseTraf  Sani
SocServ
         5          3         96   9          0          2          0          0  32
3
   Traffic     UrbInf  WaterSyst
         7        175         15
```

Shows that as for other sample years the most represented modality is UrbInf. This is due to how the Google Cloud geocode API for localization functions. It does not perform well when two streets are tangent to one another but do not physically cross one another.

Nbr of unique ZIP codes in data set = 184 after consolidation:
ZIP code 10027 absorbed ZIP code 10115
ZIP code 11101 absorbed ZIP code 11109

A Chi square test shows that there is significant association between categorical variables in the 2 way contingency table. (We reject H_0)

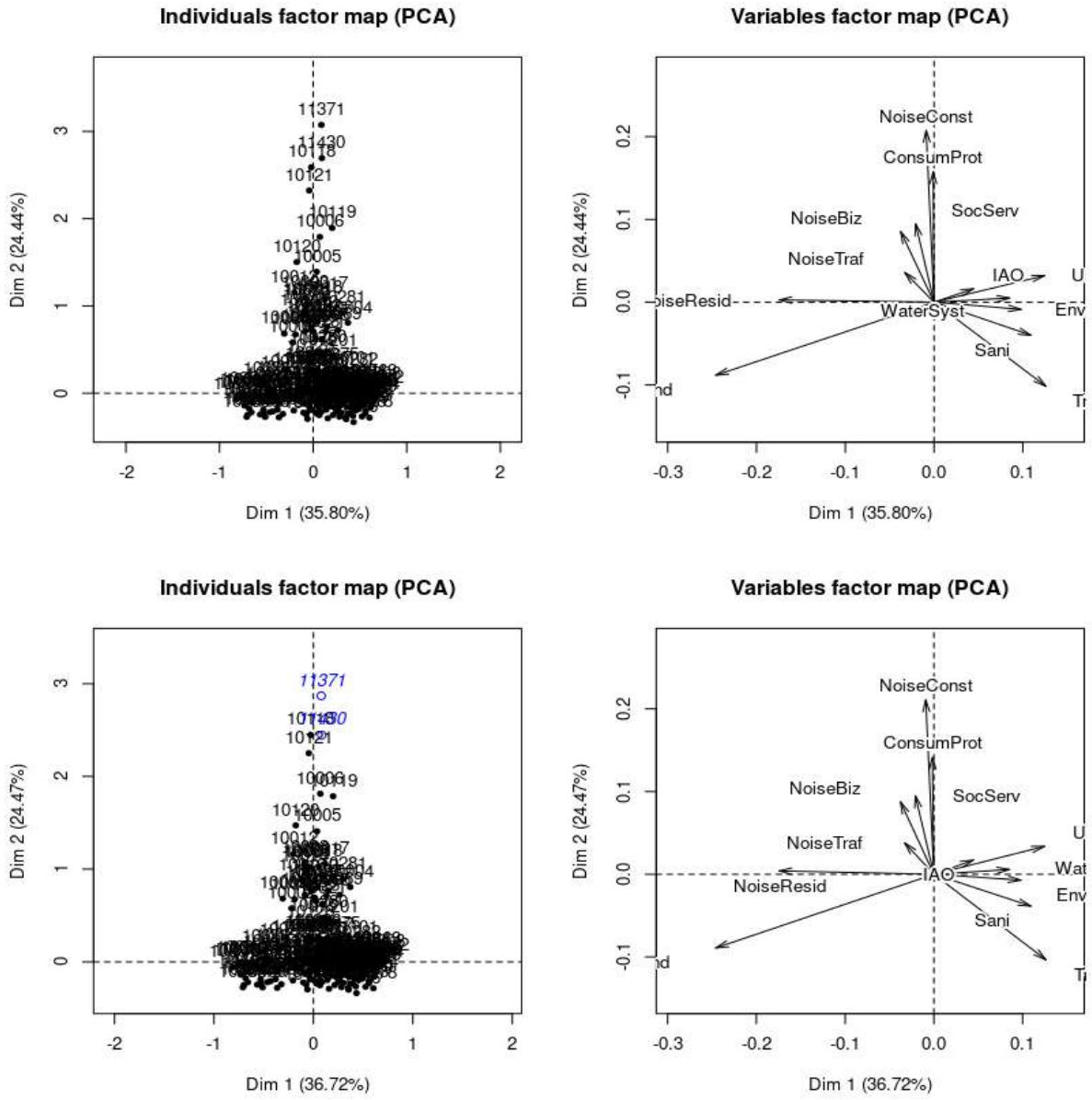
Nbr of cells with count <5: 337 . The contribution of low count cells to the computation of the Chi square statistic is negligible. No low count cells contribute 1‰ or more to the Chi square metric.

We perform a second Chi square test based on the contingency table made of low count row profiles only. Again we reject the null hyothesis as the p-value is 0.0002 for a chi square statistics of 441,

PCA on SRCs for April 2018

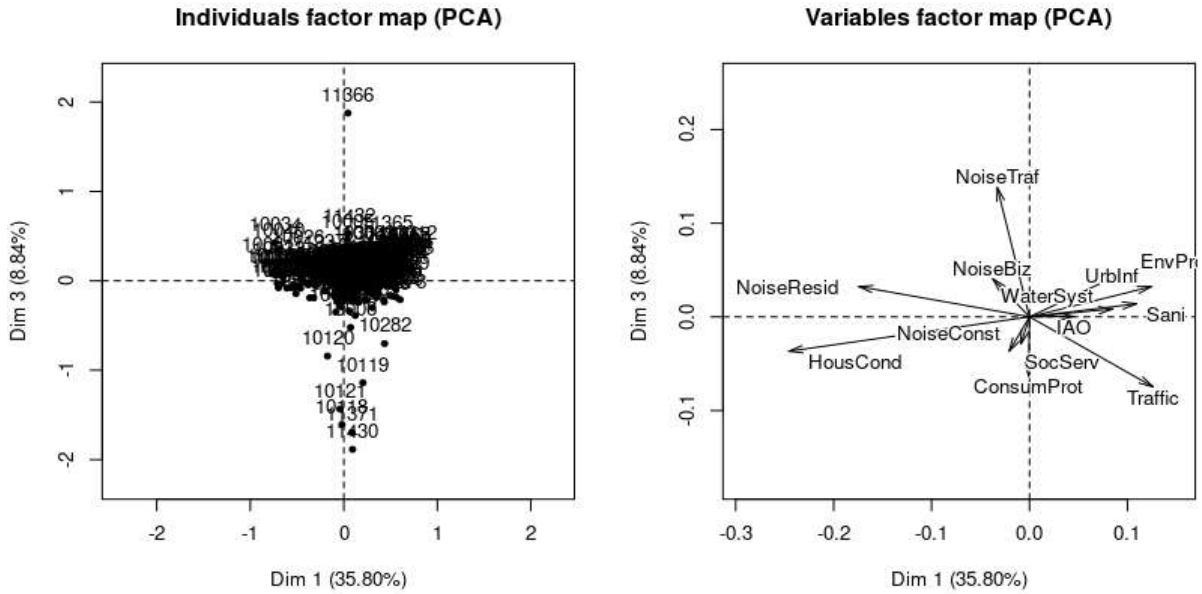
Scatter plot of scores and variables' representation in PC1-2, PC1-3 and PC2-3
The effect of ZIP code "11430" and "11371" is negligible. It is not as true an outlier as "10430" (Riverdale, the Bronx) in the April 2010 data set. Nevertheless we include them as SUP individuals.

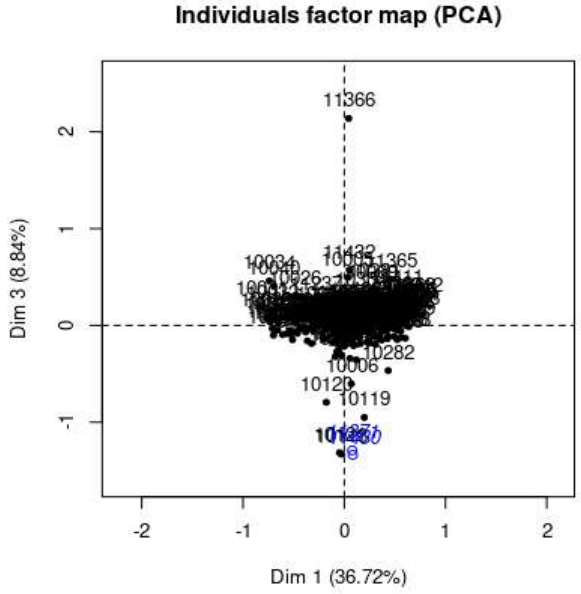
```
Eigenvalue percentage of variance cumulative percentage of variance
      (calculated considering "11430" (JFK airport) and "11371" (La Guardia
airport) as supplementary individuals.)

          eigenvalue          % of variance          cumulative % of variance
comp 1  1.561362e-01           3.580064e+01                          35.80064
comp 2  1.065764e-01           2.443702e+01                          60.23766
comp 3  3.853845e-02           8.836526e+00                          69.07419
comp 4  3.606266e-02           8.268848e+00                          77.34304
comp 5  2.621650e-02           6.011212e+00                          83.35425
comp 6  2.085334e-02           4.781486e+00                          88.13573
comp 7  1.555735e-02           3.567162e+00                          91.70289
comp 8  9.704461e-03           2.225147e+00                          93.92804
comp 9  8.686851e-03           1.991818e+00                          95.91986
comp 10 7.538626e-03           1.728540e+00                          97.64840
comp 11 6.465050e-03           1.482379e+00                          99.13078
comp 12 3.790908e-03           8.692216e-01                         100.00000


4 significant dimensions detected with criterion set at 72% of cumulated overall
variance representation.
```

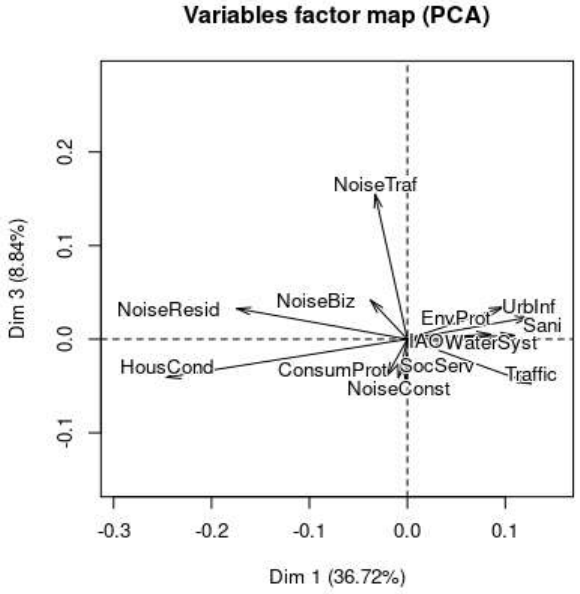


Contribution of variables to the construction of each dimension

	Dim.1	Dim.2	Dim.3	Dim.4
Sani	7.663637491	1.42659793	0.04187663	4.3902074
ConsumProt	0.001651857	19.14947109	5.09171769	1.4742624
NoiseBiz	0.904831412	7.34655210	4.61517491	2.5331514
NoiseResid	19.474381324	0.01584991	2.83865159	0.1092133
WaterSyst	4.643049276	0.03191124	0.07427844	1.0967126
Traffic	10.211814772	10.28548249	6.00289245	51.5545842
NoiseTraf	0.692606882	1.37858621	63.72175449	10.7511770
UrbInf	9.945916357	1.09751130	1.47445739	7.0005306
HousCond	38.692669750	7.61935028	4.40638568	2.8311684
SocServ	0.279180678	8.64575794	4.08063279	0.2662813
IAO	1.268239087	0.28298221	0.01425036	0.1643788
NoiseConst	0.053365702	42.67064828	4.50535636	0.2233713
EnvProt	6.168655412	0.04929903	3.13257122	17.6049612

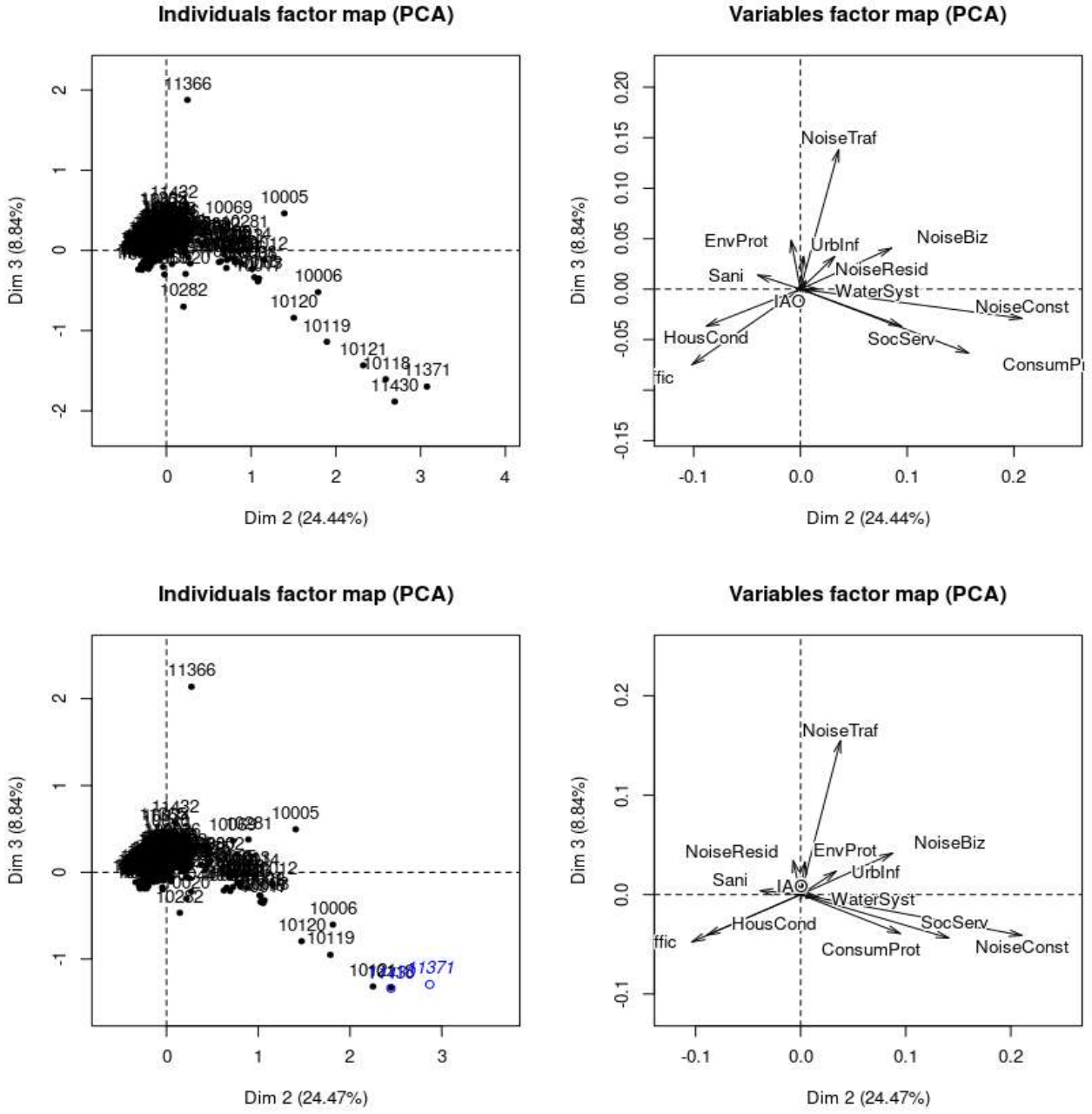






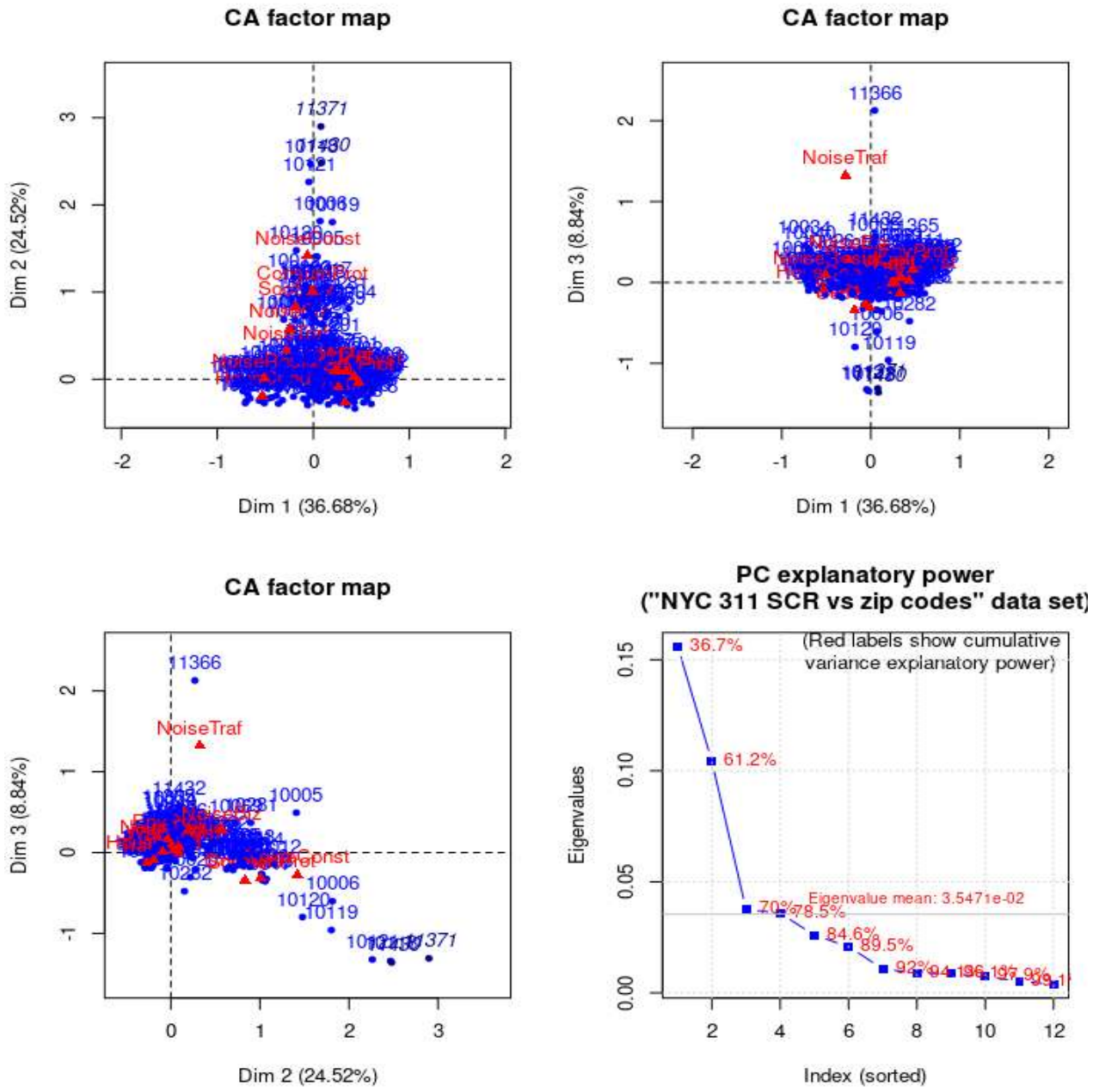
Correlation of variables with PCs:

	Dim.1	Dim.2	Dim.3	Dim.4
Sani	0.621560875	-0.21891971	0.02254436	-0.22554111
ConsumProt	-0.009573849	0.84148539	-0.26080642	0.13712123
NoiseBiz	-0.243375642	0.56611275	0.26969517	0.19522744
NoiseResid	-0.867931798	0.02021320	0.16259067	0.03116079
WaterSyst	0.664396733	0.04496406	0.04123282	-0.15480653
Traffic	0.574920985	-0.47101736	-0.21628320	0.61930893
NoiseTraf	-0.175736253	0.20239641	0.82708023	0.33194258
UrbInf	0.739491747	0.20053183	0.13970529	-0.29743592
HousCond	-0.898108992	-0.32534306	-0.14871055	-0.11647013
SocServ	-0.161429826	0.73334868	-0.30282427	0.07558375
IAO	0.558425048	0.21533358	-0.02904444	-0.09638383
NoiseConst	-0.039967119	0.92258042	-0.18018645	0.03920148
EnvProt	0.637931760	-0.04655497	0.22305669	-0.51667031



Generally speaking as for other time periods (April 2010 and 2014):

- "NoiseTraf" and "HousCond" are nearly orthogonal.
- etc.



Contributions to the construction of principal directions > 10%

	Dim 1	Dim 2	Dim 3
ConsumProt	0.001714527	19.42535316	5.197787
NoiseResid	19.473149010	0.01518322	2.832599
Traffic	10.215163807	10.22457239	6.365455
NoiseTraf	0.692681141	1.36325022	63.191799
HousCond	38.688309435	7.59334245	4.370598
NoiseConst	0.053676847	42.54405996	4.434335

Inertia explanatory power for all dimensions and for the significant dimensions

	iep_alldim	iep_sigdim
Sani	7.3	4.5
ConsumProt	6.7	6.8
NoiseBiz	5.6	3.5
NoiseResid	**9.5**	**9.4**
WaterSyst	3.9	2.3
Traffic	**11.3**	**14.2**
NoiseTraf	**8.2**	**9.1**
UrbInf	6.7	5.9
HousCond	**17.6**	**21.3**
SocServ	3.9	3.3
IAO	1.5	0.7
NoiseConst	**12.3**	**13.8**
EnvProt	5.6	5.1

Quality of representation of col profiles with biggest contrib to PC formation

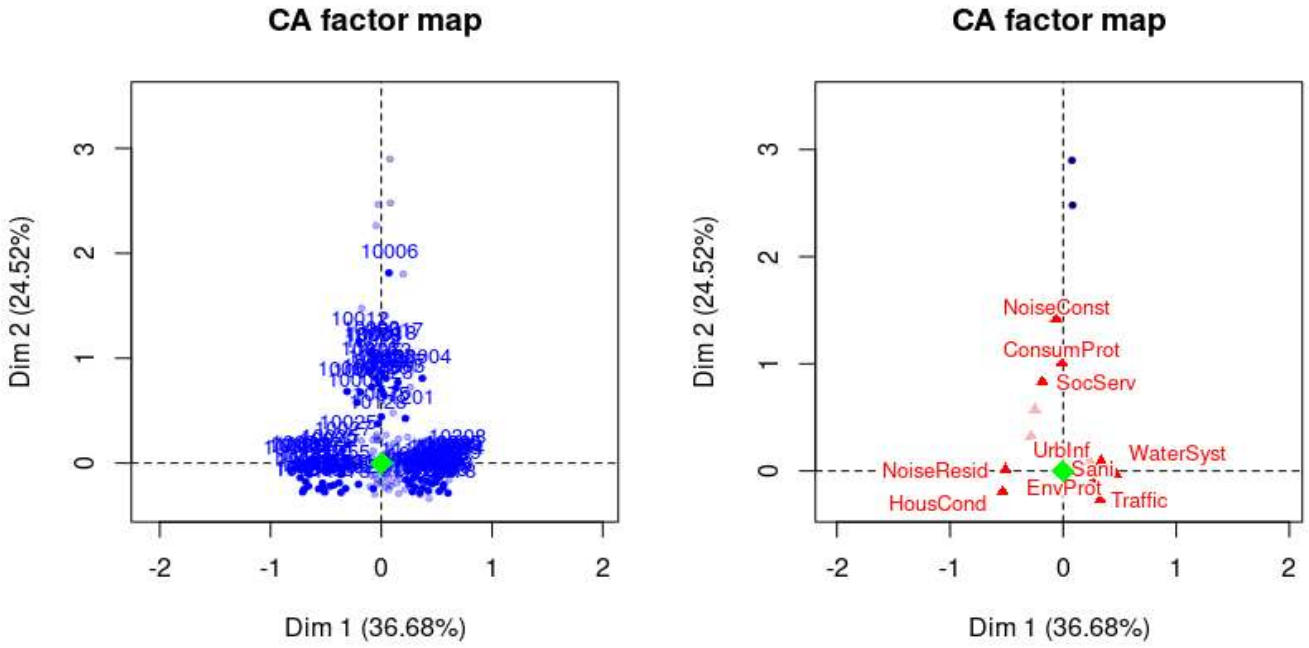
	Dim 1	Dim 2	Dim 3
ConsumProt	0.00009377674	0.7101053632	0.06849580
NoiseResid	0.75330020880	0.0003925537	0.02640047
Traffic	0.33060387467	0.2211622162	0.04963486
NoiseTraf	0.03088827521	0.0406292725	0.67891557
HousCond	0.80653471136	0.1057984906	0.02195225

Most important individual contributors to the construction of PCs

	Dim 1	Dim 2	Dim 3	
10001	0.0	5.1	1.0	
10002	0.2	2.6	0.4	
10003	**0.0**	**8.7**	**2.2**	**--- max contr to construct of Dim 2**
10006	0.0	2.4	0.7	
10009	0.4	3.0	0.1	
10011	0.0	3.9	0.0	
10012	0.1	6.0	0.3	
10013	0.0	3.1	0.0	
10014	0.0	4.8	0.0	
10016	0.0	4.1	0.5	
10017	0.0	2.2	0.7	
10018	0.0	2.5	0.8	
10019	0.0	4.3	0.3	
10022	0.0	2.5	0.3	
10023	0.0	2.3	0.5	
10031	**4.7**	**0.0**	**1.2**	**--- max contr to construct of Dim 1**
10032	2.1	0.0	0.4	
10033	3.0	0.0	0.1	
10034	2.7	0.0	4.3	
10036	0.0	2.3	0.6	
10038	0.0	2.1	0.4	
10040	2.1	0.0	2.9	
10312	2.4	0.0	1.1	
10314	3.1	0.0	0.6	

```
10452   3.2   0.5   0.1
10453   3.7   0.8   0.0
10458   3.7   0.7   0.3
10467   2.6   0.7   0.3
10468   3.2   0.3   0.1
11226   2.2   0.7   0.3
11365   0.6   0.0   2.5
11366   0.0   0.3  45.7   --- max contr to construct of Dim 3
11432   0.0   0.1   5.8
```

Results for CA:

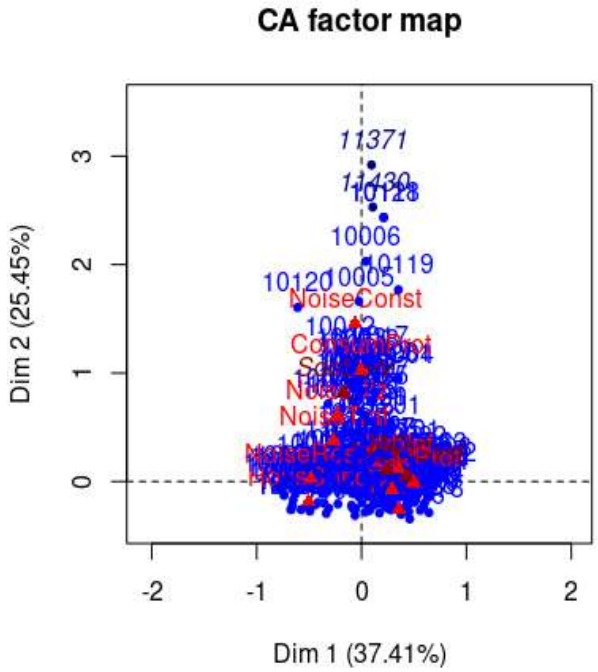


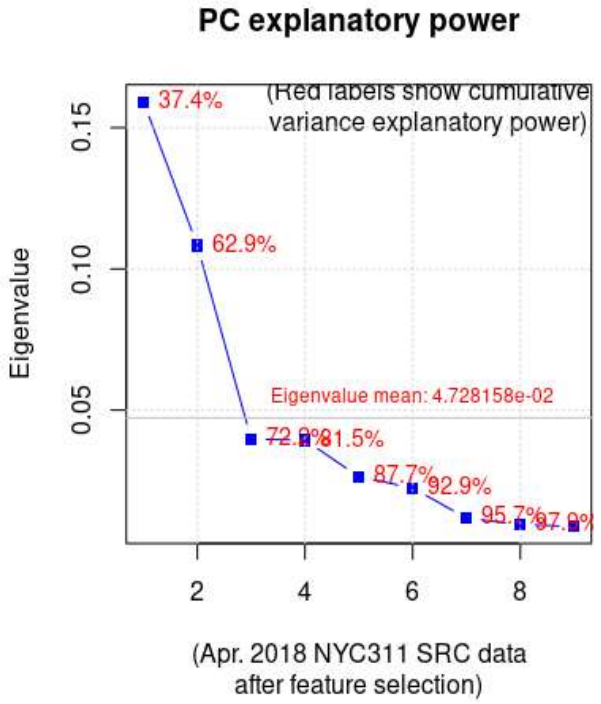
Left row profiles with cos² > 0.6 – Right col profiles with cos² > 0.4

After feature selection (i.e. setting “WaterSyst”, “SocServ” and “IAO” as SUP variables):

```
      eigenvalue percentage of variance cumulative percentage of variance
dim 1 0.159204045             37.412748                          37.41275
dim 2 0.108277422             25.445056                          62.85780
dim 3 0.039776913              9.347524                          72.20533
dim 4 0.039391240              9.256891                          81.46222
dim 5 0.026394997              6.202791                          87.66501
dim 6 0.022423225              5.269429                          92.93444
dim 7 0.011761538              2.763947                          95.69839
dim 8 0.009485556              2.229094                          97.92748
dim 9 0.008819280              2.072520                         100.00000
```

→ 2 significant dimensions only





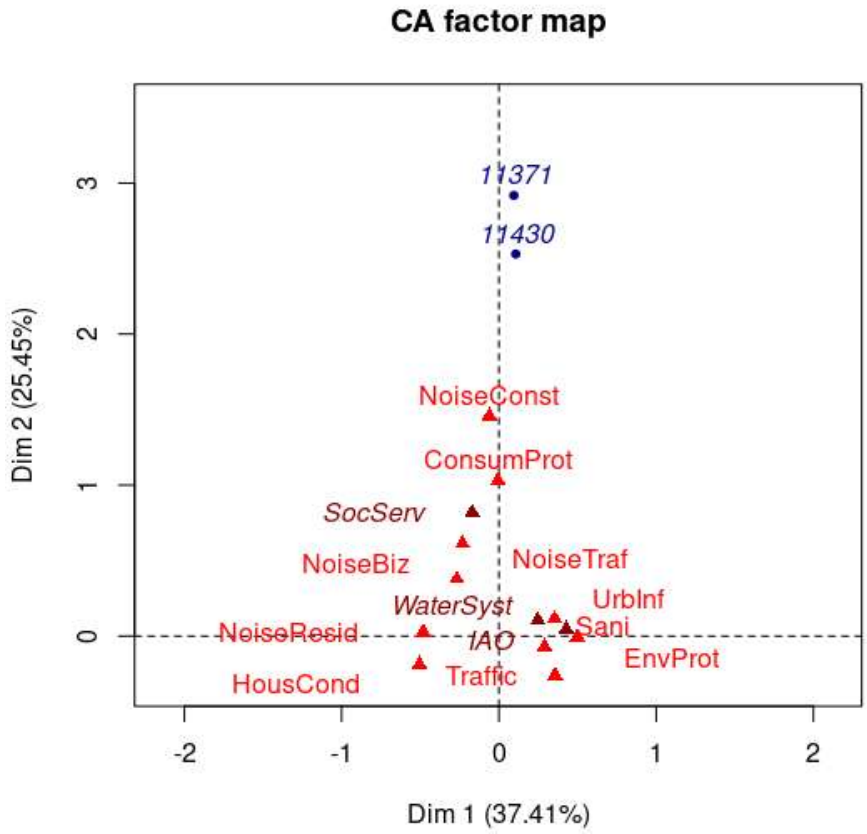
Eigenvalues

	Dim.1	**Dim.2**	**Dim.3**	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9
Variance	0.159	0.108	0.040	0.039	0.026	0.022	0.012	0.009	0.009
% of var.	37.413	25.445	9.348	9.257	6.203	5.269	2.764	2.229	2.073
Cumulative % of var.	37.413	62.858	72.205	81.462	87.665	92.934	95.698	97.927	100.000

The chi square of independence between the two variables (ZIPs + SRCs) is equal to 45880.48 (p-value = 0).

Columns

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
Sani	34.824	0.290	10.179	0.465	-0.072	0.913	0.028	-0.012	0.069	0.001
ConsumProt	34.158	-0.006	0.000	0.000	1.030	21.429	0.679	-0.326	5.834	0.068
NoiseBiz	25.911	-0.232	0.851	0.052	0.611	8.703	0.364	0.228	3.285	0.050
NoiseResid	40.378	-0.481	18.755	0.739	0.026	0.082	0.002	0.077	1.902	0.019
Traffic	53.932	0.357	12.933	0.382	-0.260	10.068	0.202	-0.062	1.555	0.011
NoiseTraf	38.212	-0.267	0.669	0.028	0.378	1.970	0.056	1.402	73.839	0.769
UrbInf	35.443	0.354	11.944	0.536	0.115	1.843	0.056	0.042	0.681	0.008
HousCond	74.312	-0.505	37.215	0.797	-0.188	7.535	0.110	-0.090	4.682	0.025
NoiseConst	60.505	-0.061	0.056	0.001	1.458	47.454	0.849	-0.332	6.695	0.044
EnvProt	27.858	0.498	7.398	0.423	-0.008	0.003	0.000	0.110	1.457	0.021



Active variable are in red, while supplementary ones are in dark red. The blue point is the supplementary individual(s) ear-marked as outlier(s).

Contributions (> 10%) to the construction of the 3 first PCs after feature extraction/selection:

	Dim 1	Dim 2	Dim 3
Sani	10.1790494853	0.91300192	0.06893129
ConsumProt	0.0004277188	21.42883722	5.83449066
NoiseResid	18.7545648111	0.08226928	1.90219037
Traffic	12.9330401060	10.06849957	1.55523628
NoiseTraf	0.6692209145	1.96950630	73.83854163
UrbInf	11.9437459610	1.84344646	0.68125174
HousCond	37.2151691227	7.53488560	4.68218845
NoiseConst	0.0557624983	47.45380132	6.69527668

Quality of representations within the former group (above):

	Dim 1	Dim 2	Dim 3
Sani	0.46535969758	0.028388147	0.0007873629
ConsumProt	0.00001993502	0.679267301	0.0679419843
NoiseResid	0.73946893231	0.002206147	0.0187388918
Traffic	0.38177384513	0.202140862	0.0114704092
NoiseTraf	0.02788176901	0.055807427	0.7686186420
UrbInf	0.53648739344	0.056316195	0.0076454592
HousCond	0.79728830113	0.109788251	0.0250623198
NoiseConst	0.00146724118	0.849208050	0.0440154147

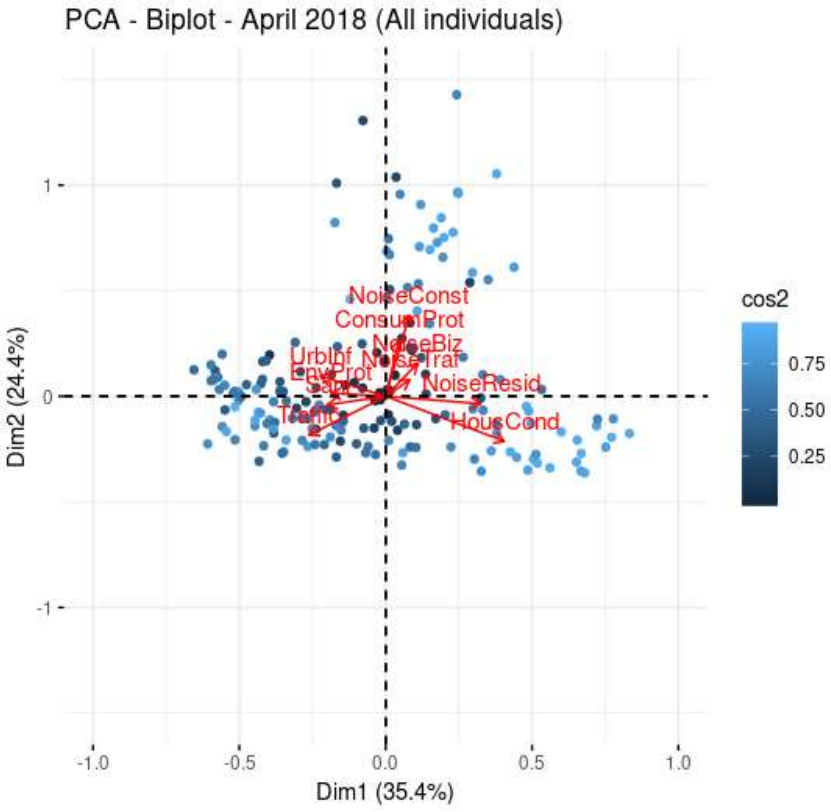
Individual contributions for which at least one component is greater than 20% in PC1-2-3

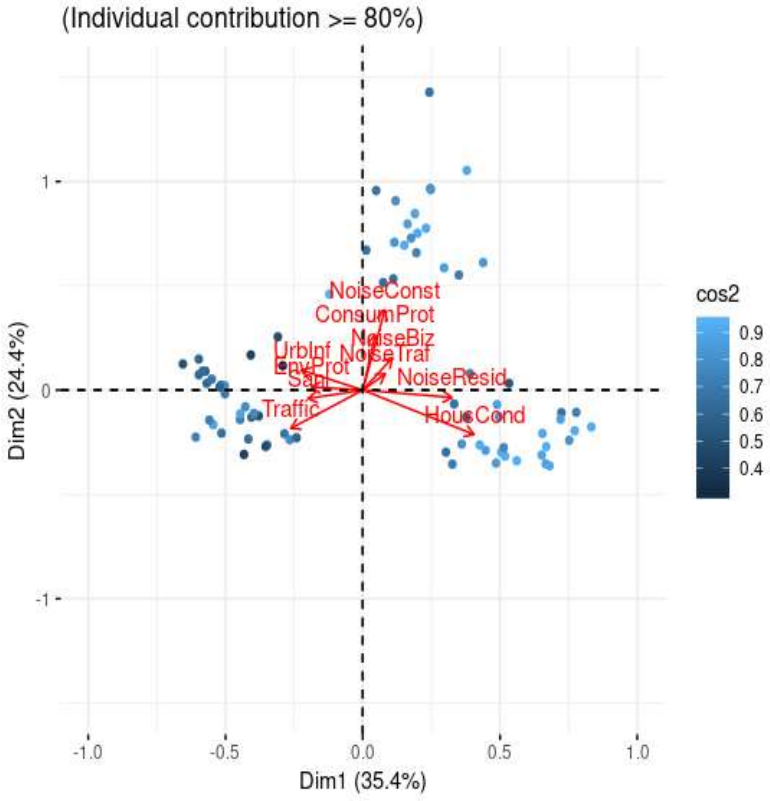
	Dim 1	Dim 2	Dim 3	
10001	0.0	4.4	0.7	
10002	0.3	3.0	0.2	
10003	**0.0**	**8.5**	**2.3**	**---- Dim2**
10005	0.0	2.4	0.7	
10006	0.0	2.7	0.9	
10009	0.4	3.2	0.2	
10011	0.0	3.8	0.1	
10012	0.1	6.0	0.3	
10013	0.0	3.2	0.0	
10014	0.0	5.4	0.2	
10016	0.0	3.8	0.5	
10017	0.0	2.1	0.6	
10018	0.0	2.4	0.7	
10019	0.0	4.4	0.3	
10022	0.0	2.4	0.3	
10031	**4.6**	**0.0**	**0.9**	**----- Dim1**
10032	2.1	0.0	0.4	
10033	3.0	0.0	0.1	
10034	2.7	0.0	4.3	
10040	2.0	0.0	2.9	
10312	2.6	0.0	0.5	
10314	2.9	0.0	0.2	
10452	3.3	0.5	0.1	
10453	3.5	0.8	0.0	
10456	2.1	0.6	0.2	
10458	3.5	0.8	0.4	
10467	2.4	0.6	0.4	
10468	3.1	0.3	0.1	
11226	2.1	0.7	0.2	
11365	0.7	0.0	2.6	
11366	**0.0**	**0.4**	**55.1**	**---- Dim3**
11432	0.0	0.1	6.7	

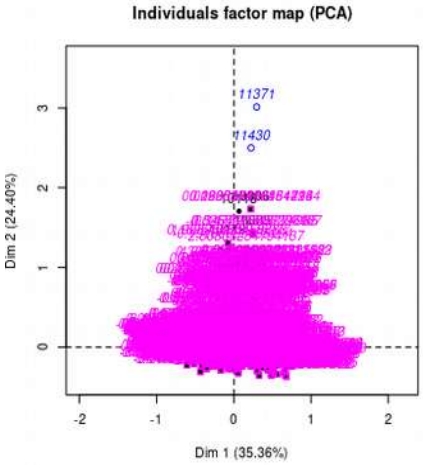
Not changed from before. Good sign.

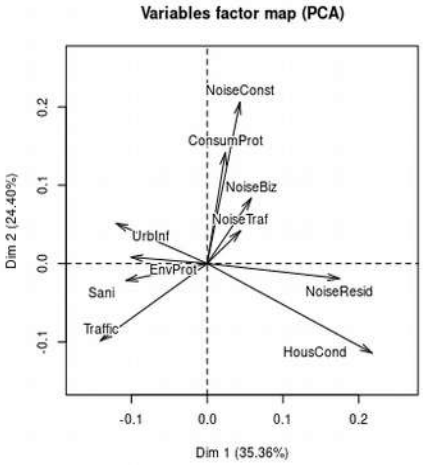
Inertia Explanatory Power for factors (i.e. variable's modalities)

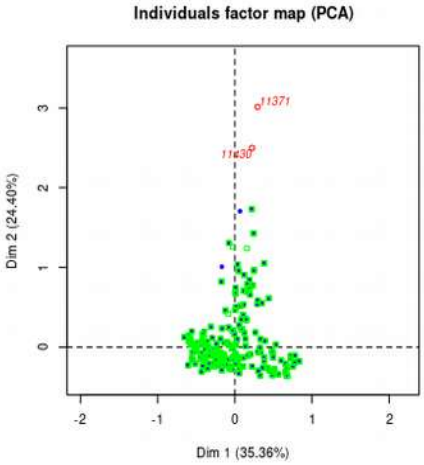
	iep_alldim	iep_sigdim
	iep_alldim	iep_sigdim
Sani	8.2	5.6
ConsumProt	8.0	8.3
NoiseBiz	6.1	3.9
NoiseResid	**9.5**	**10.0**
Traffic	**12.7**	**10.5**
NoiseTraf	**9.0**	**10.6**
UrbInf	8.3	6.9
HousCond	**17.5**	**22.5**
NoiseConst	**14.2**	**17.6**
EnvProt	6.5	4.0

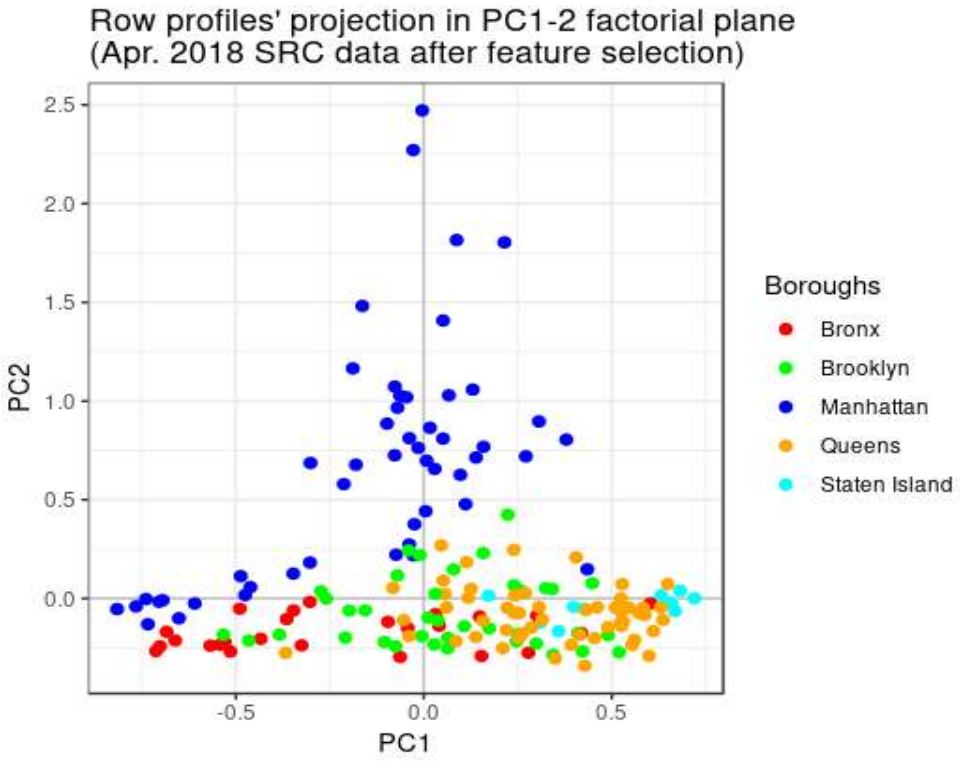

PCA - Biplot - April 2018 (All individuals)
Dim2 (24.4%)
Dim1 (35.4%)
cos2
0.75
0.50
0.25
NoiseConst
ConsumProt
NoiseBiz
UrblInf
NoiseTraf
EnvProt
Safety
NoiseResid
Traffic
HousCond
1
0
-1
-1.0
-0.5
0.0
0.5
1.0


(Individual contribution >= 80%)
Dim2 (24.4%)
Dim1 (35.4%)
cos2
0.9
0.8
0.7
0.6
0.5
0.4
NoiseConst
ConsumProt
NoiseBiz
UrblInf
NoiseTraf
EnvProt
Safety
NoiseResid
Traffic
HousCond
1
0
-1
-1.0
-0.5
0.0
0.5
1.0


Individuals factor map (PCA)
11371
11430
Dim 2 (24.40%)
Dim 1 (35.36%)

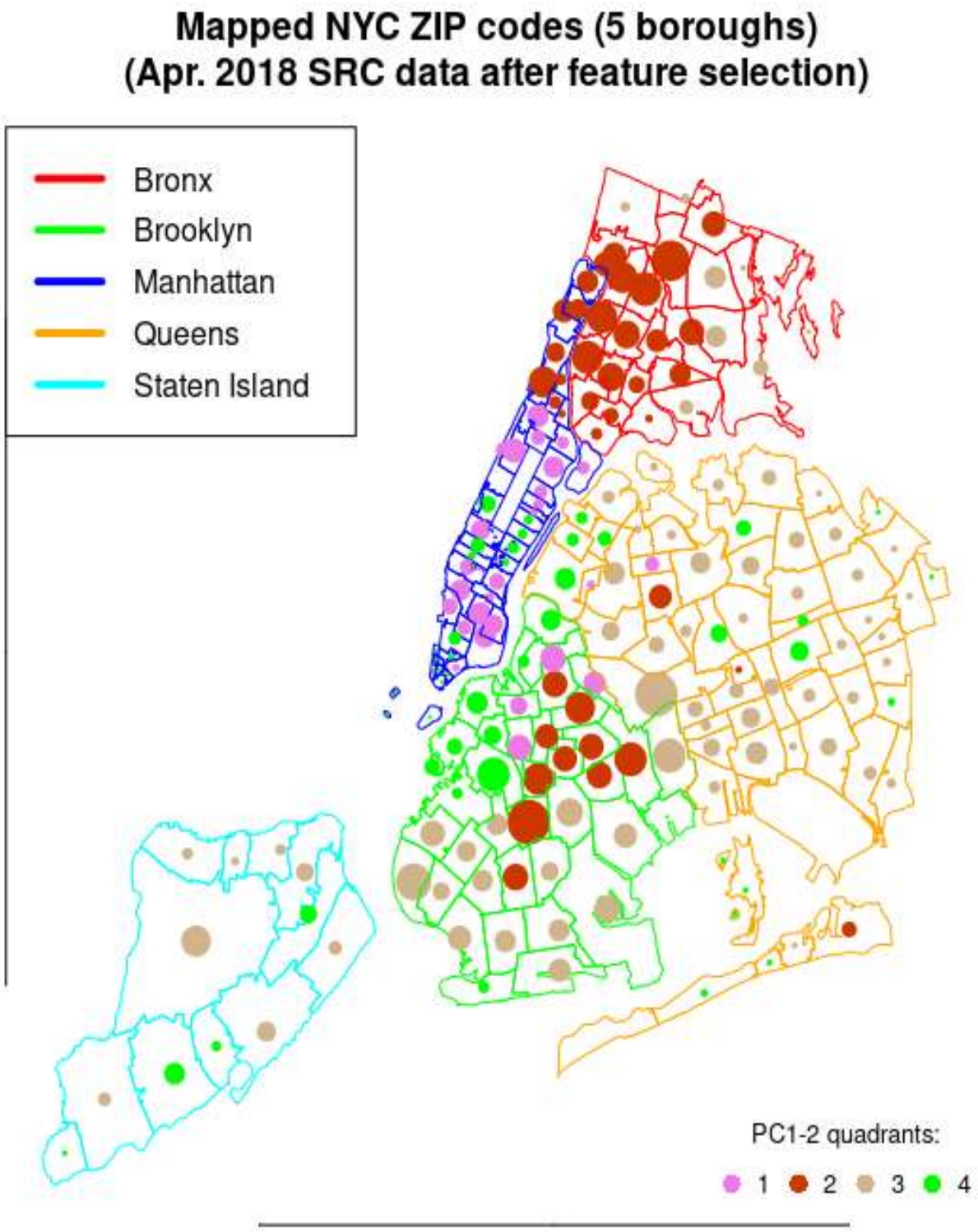

Variables factor map (PCA)
NoiseConst
ConsumProt
NoiseBiz
NoiseTraf
UrbInf
EnvProt
Sani
Traffic
NoiseResid
HousCond
Dim 2 (24.40%)
Dim 1 (35.36%)


Individuals factor map (PCA)
11371
11430
Dim 2 (24.40%)
Dim 1 (35.36%)

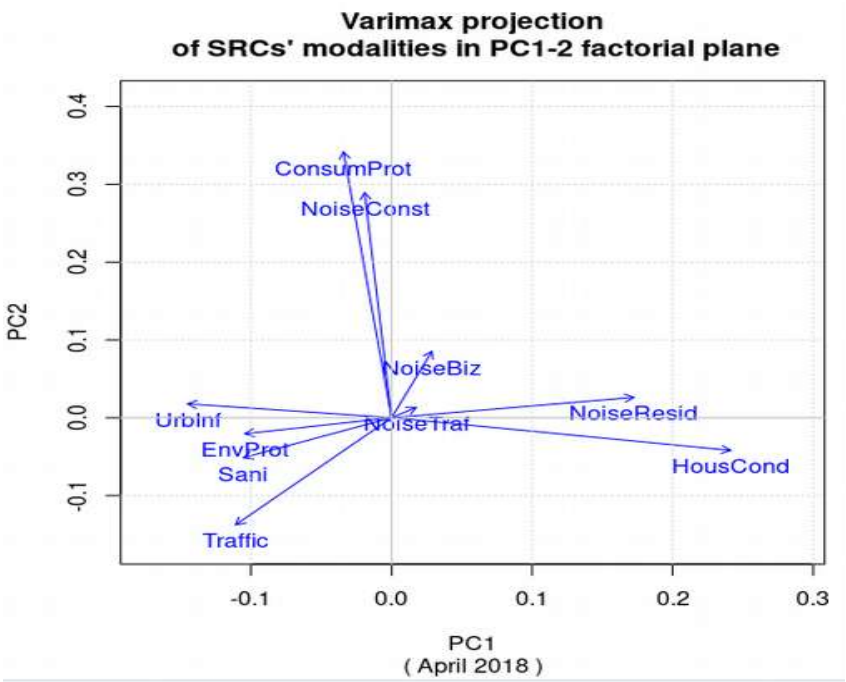

Row profiles' projection in PC1-2 factorial plane
(Apr. 2018 SRC data after feature selection)
PC2
PC1
Boroughs
Bronx
Brooklyn
Manhattan
Queens
Staten Island

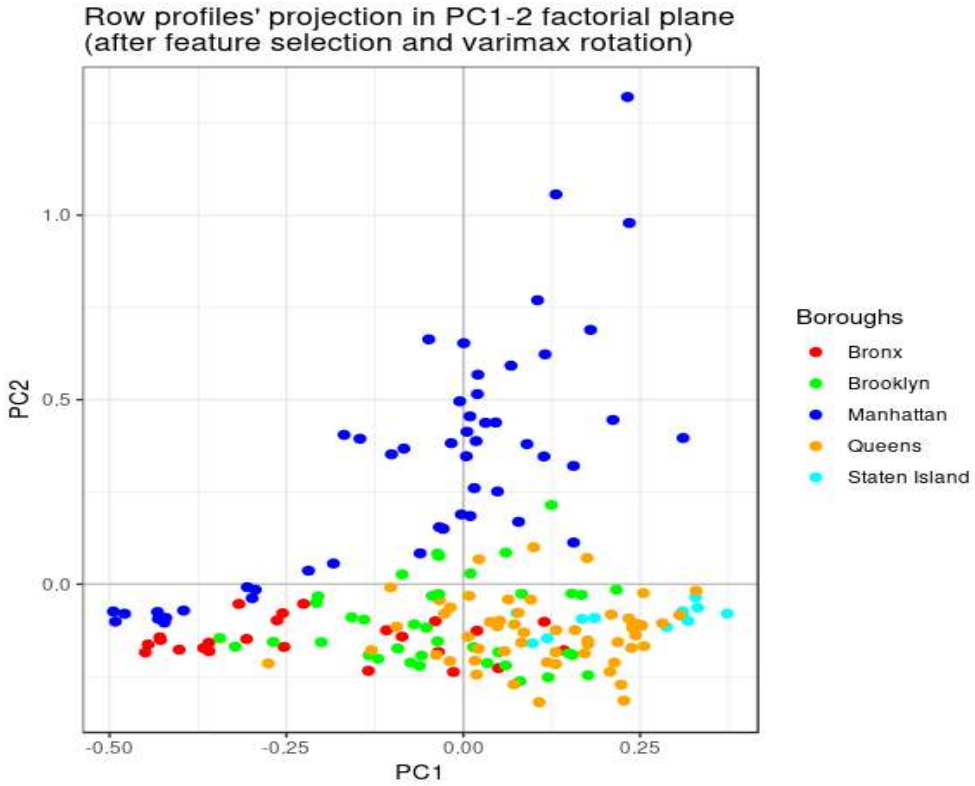
Analysis of ZIP codes' (observations') contributions to inertia per borough:

```
Borough:  Manhattan
 Number of ZIP codes:  50
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   34.0    27.7

Borough:  Staten Island
 Number of ZIP codes:  12
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
    9.2     4.9

Borough:  Bronx
 Number of ZIP codes:  25
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   13.1    10.8

Borough:  Queens
 Number of ZIP codes:  57
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   28.1    19.9

Borough:  Brooklyn
 Number of ZIP codes:  38
Borough's ZIPs' % contribution to inertia (overall and in PC1-2 factorial plane):
All_dim   PC1-2
   14.9     8.3
```

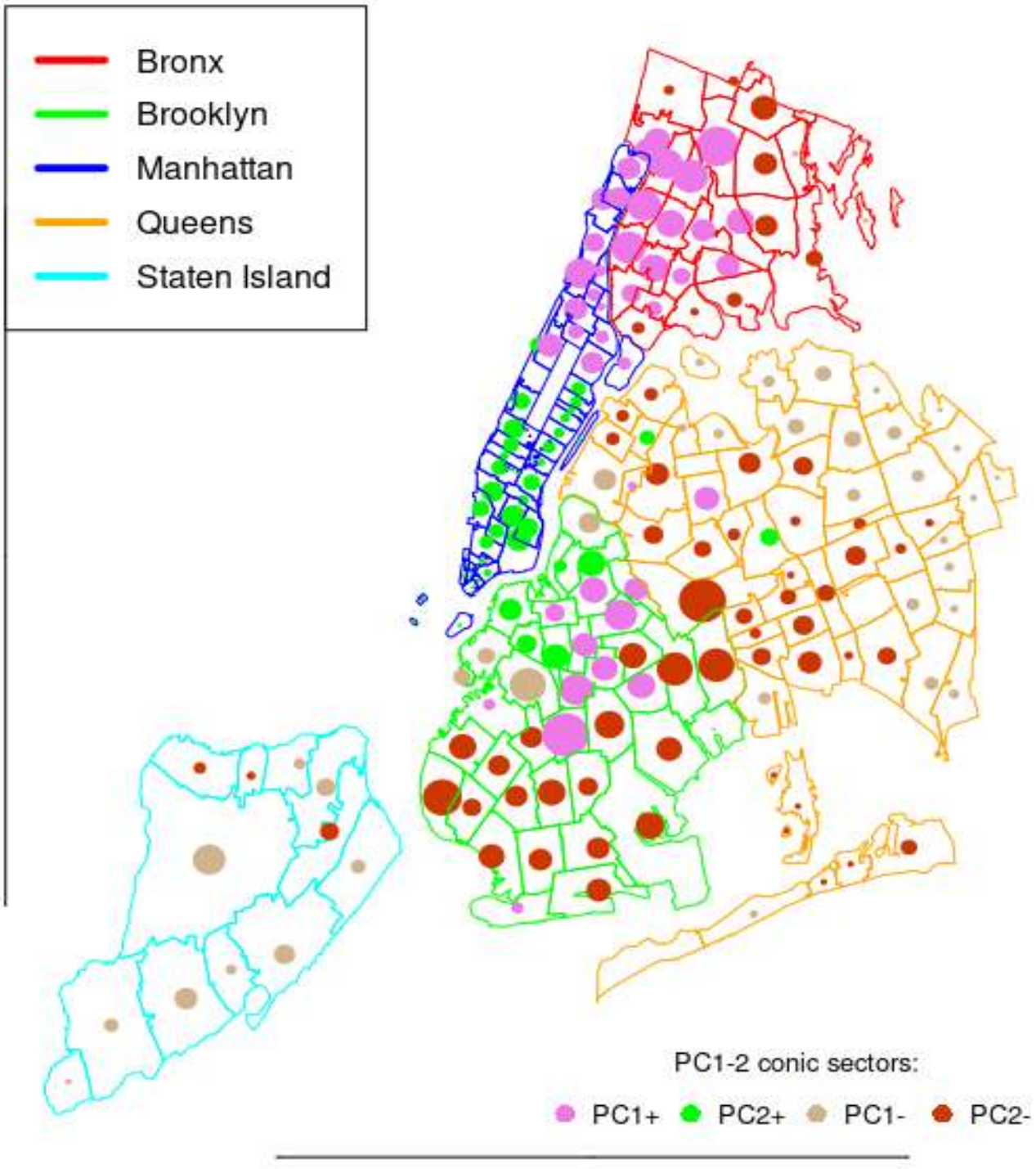

Mapped NYC ZIP codes (5 boroughs)
(Apr. 2018 SRC data after feature selection)
Bronx
Brooklyn
Manhattan
Queens
Staten Island
PC1-2 quadrants:
1
2
3
4

Analysis of latent features following varimax treatment:





Mapped NYC ZIP codes (5 boroughs) - Apr. 2018
(after feature selection and varimax rotation)



MCA with crime data:

Binning the 17 categorical variables brings some change when compared to reference year 2014, in particular in the distribution and frequency of "HousCond" related SRCs.

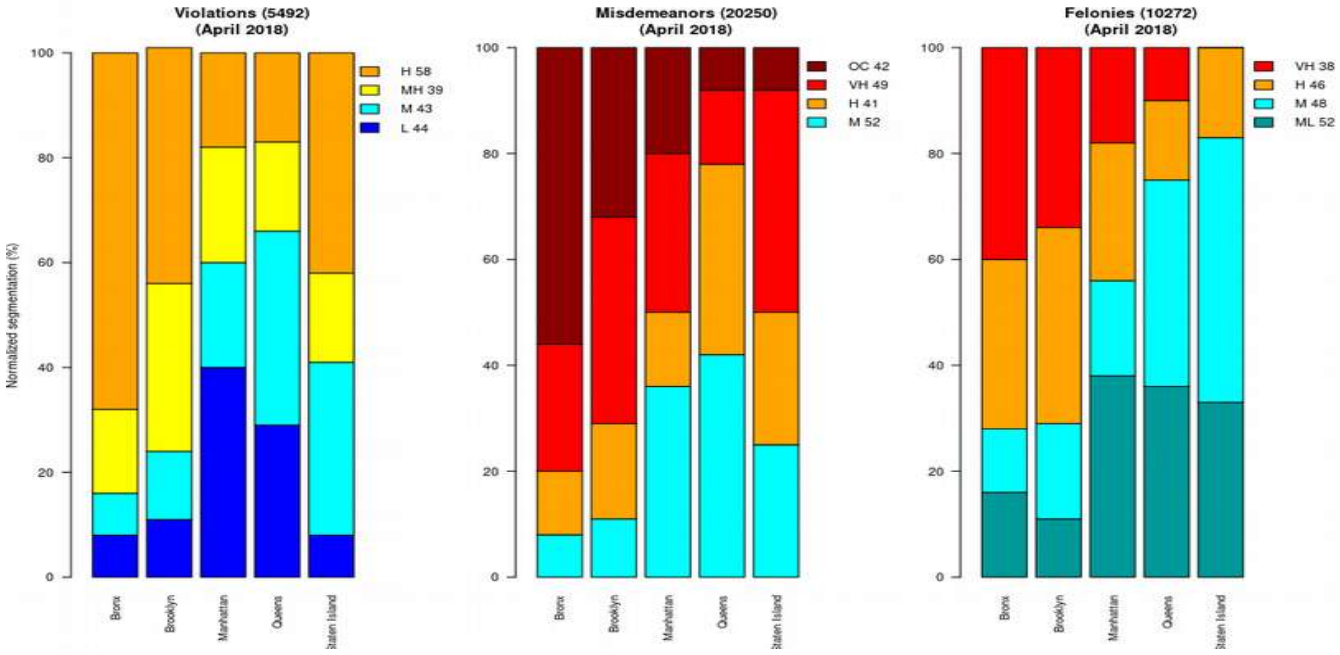
```
HousCond:
   below or equal to 56 SRCs - bin count: 48
   between 57 and 137 SRCs - bin count: 45
   between 138 and 298 SRCs- bin count: 46
   above 298 SRCs - bin count: 46
Sani:
   below or equal to 18 SRCs - bin count: 17
   between 19 and 31 SRCs - bin count: 5
   between 32 and 54 SRCs- bin count: 5
   above 54 SRCs - bin count: 158
NoiseResid:
   below or equal to 29 SRCs - bin count: 44
   between 30 and 61 SRCs - bin count: 29
   between 62 and 123 SRCs- bin count: 48
   above 123 SRCs - bin count: 64
NoiseConst:
   below or equal to 1 SRCs - bin count: 28
   between 2 and 5 SRCs - bin count: 39
   between 6 and 20 SRCs- bin count: 63
   above 20 SRCs - bin count: 55
NoiseBiz:
   below or equal to 3 SRCs - bin count: 41
   between 4 and 9 SRCs - bin count: 42
   between 10 and 27 SRCs- bin count: 50
   above 27 SRCs - bin count: 52
UrbInf:
   below or equal to 33 SRCs - bin count: 13
   between 34 and 55 SRCs - bin count: 6
   between 56 and 87 SRCs- bin count: 23
   above 87 SRCs - bin count: 143
Traffic:
   below or equal to 24 SRCs - bin count: 21
   between 25 and 54 SRCs - bin count: 25
   between 55 and 87 SRCs- bin count: 26
   above 87 SRCs - bin count: 113
NoiseTraf:
   below or equal to 4 SRCs - bin count: 63
   between 5 and 11 SRCs - bin count: 46
   between 12 and 23 SRCs- bin count: 44
   above 23 SRCs - bin count: 32
WaterSyst:
   below or equal to 19 SRCs - bin count: 39
   between 20 and 29 SRCs - bin count: 31
   between 30 and 44 SRCs- bin count: 44
   above 44 SRCs - bin count: 71
ConsumProt:
   below or equal to 5 SRCs - bin count: 38
   between 6 and 13 SRCs - bin count: 46
   between 14 and 23 SRCs- bin count: 53
   above 23 SRCs - bin count: 48
SocServ:
   below or equal to 2 SRCs - bin count: 34
   between 3 and 6 SRCs - bin count: 38
   between 7 and 11 SRCs- bin count: 24
   above 11 SRCs - bin count: 89
```

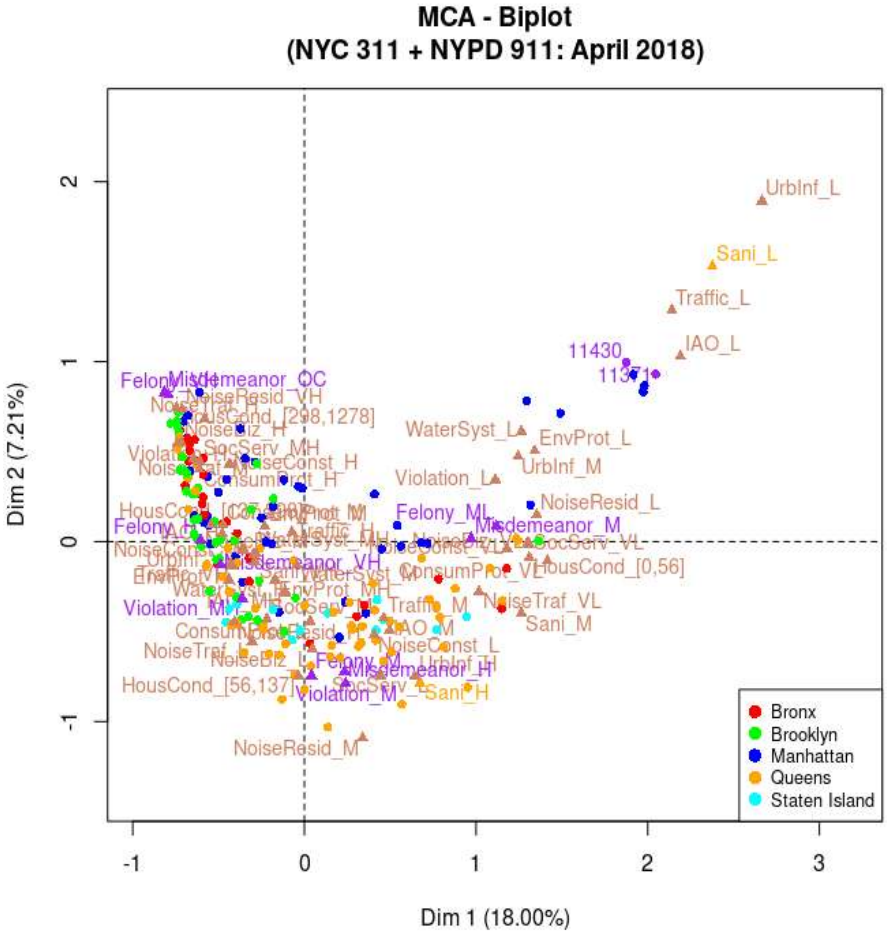
```
IAO:
   below or equal to 14 SRCs - bin count: 20
   between 15 and 23 SRCs - bin count: 27
   between 24 and 33 SRCs- bin count: 35
   above 33 SRCs - bin count: 103
EnvProt:
   below or equal to 16 SRCs - bin count: 37
   between 17 and 26 SRCs - bin count: 30
   between 27 and 41 SRCs- bin count: 45
   above 41 SRCs - bin count: 73
Violation:
   below or equal to 7 SRCs - bin count: 45
   between 8 and 20 SRCs - bin count: 43
   between 21 and 38 SRCs- bin count: 39
   above 38 SRCs - bin count: 58
Misdemeanor:
   below or equal to 33 SRCs - bin count: 53
   between 34 and 87 SRCs - bin count: 41
   between 88 and 178 SRCs- bin count: 49
   above 178 SRCs - bin count: 42
Felony:
   below or equal to 17 SRCs - bin count: 53
   between 18 and 45 SRCs - bin count: 48
   between 46 and 91 SRCs- bin count: 46
   above 91 SRCs - bin count: 38
```

	Violation			
Borough	L	M	MH	H
Bronx	2	2	4	17
Brooklyn	4	5	12	17
Manhattan	20	10	11	9
Queens	17	22	10	10
Staten Island	1	4	2	5

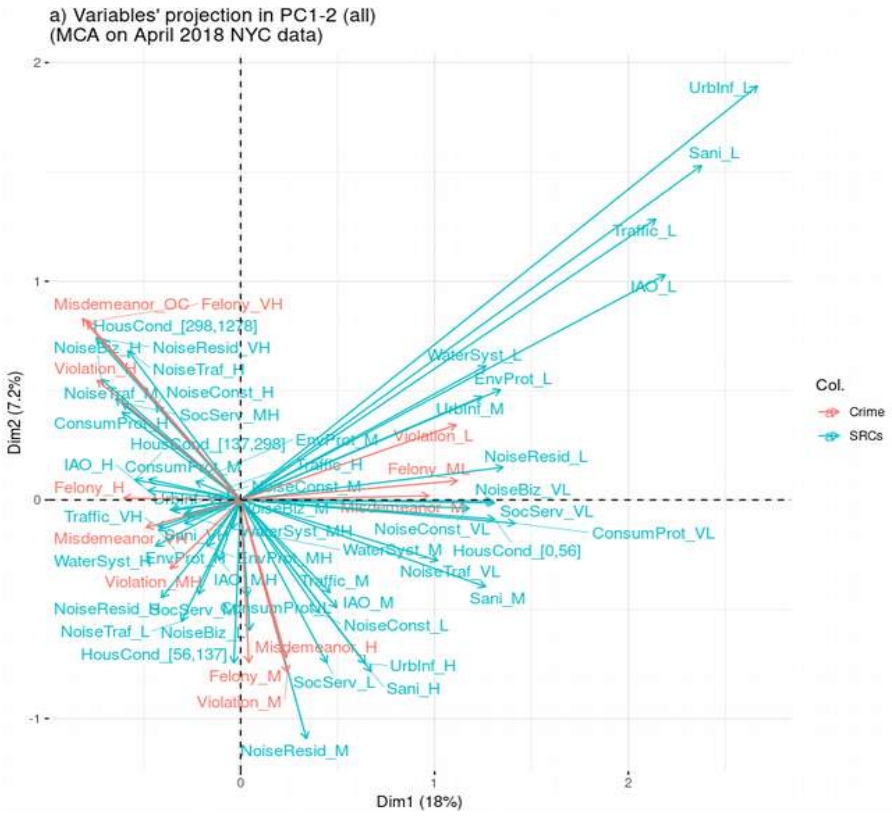
	Misdemeanor			
Borough	M	H	VH	OC
Bronx	2	3	6	14
Brooklyn	4	7	15	12
Manhattan	18	7	15	10
Queens	25	21	8	5
Staten Island	3	3	5	1

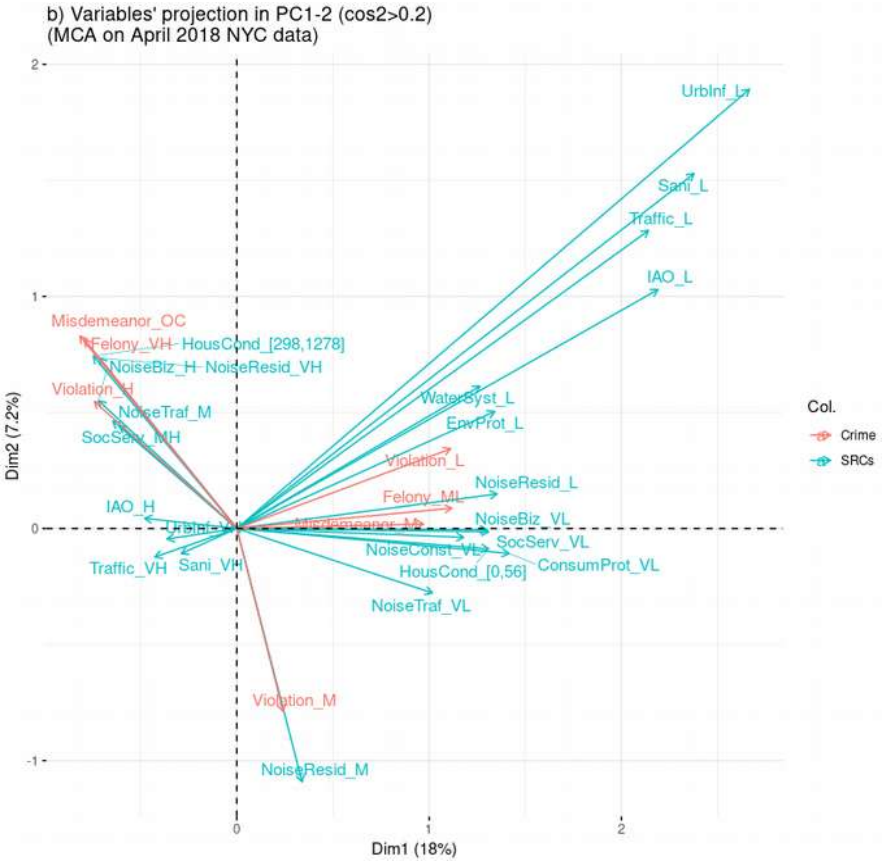
	Felony			
Borough	ML	M	H	VH
Bronx	4	3	8	10
Brooklyn	4	7	14	13
Manhattan	19	9	13	9
Queens	21	23	9	6
Staten Island	4	6	2	0

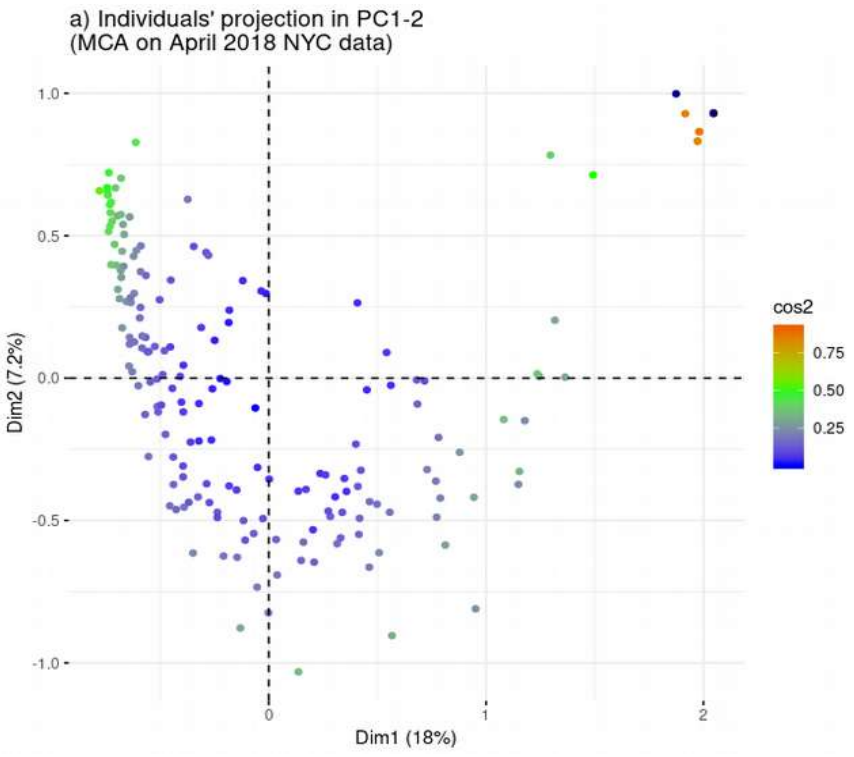


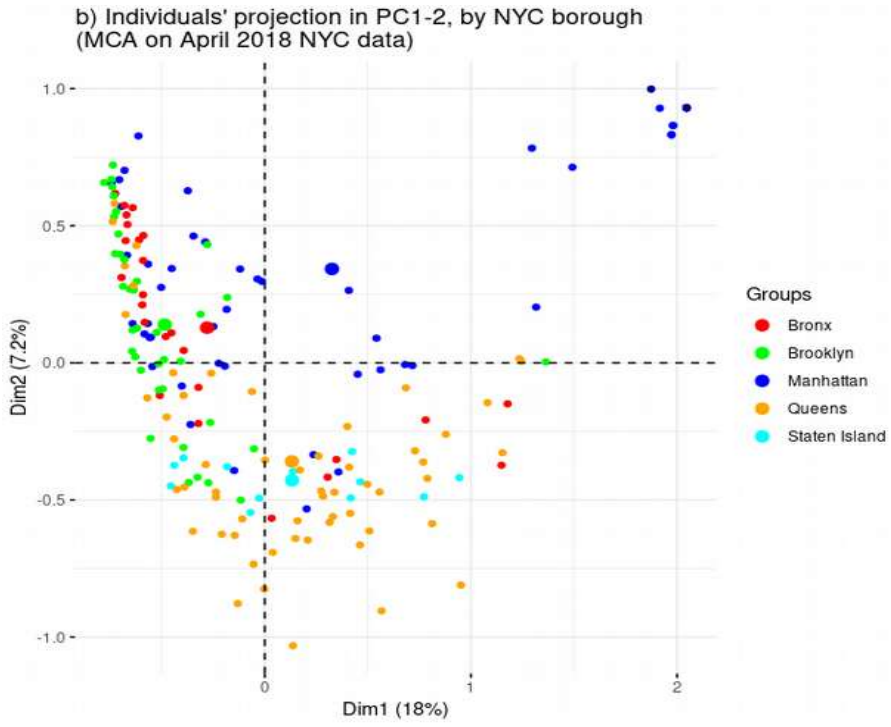


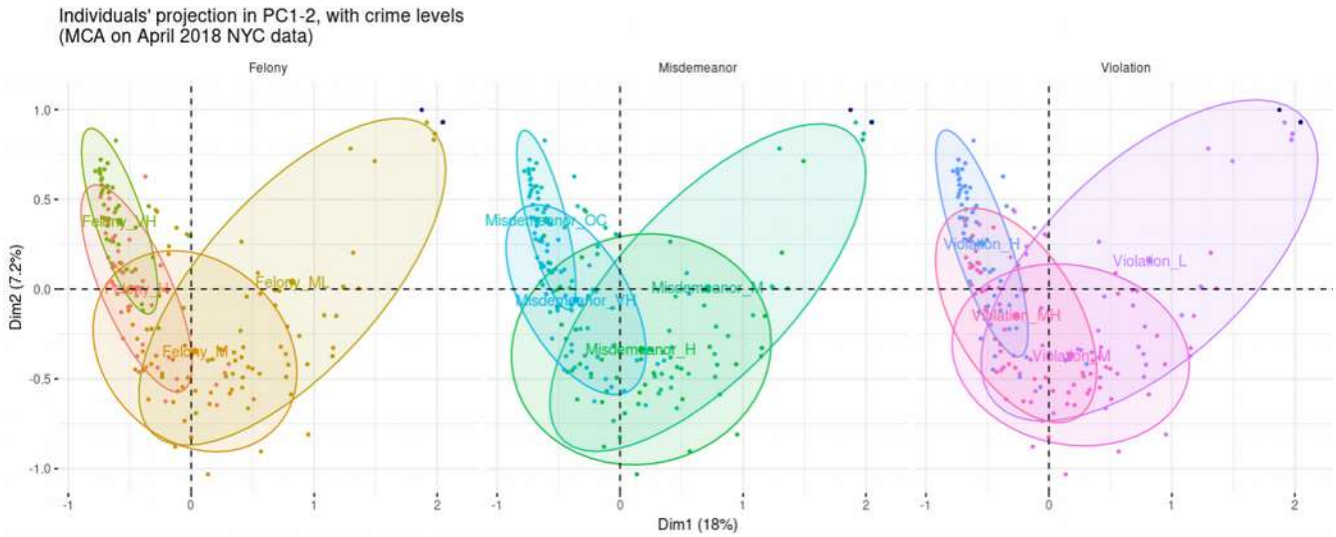
- Crime factors and supplementary individuals in purple
- Other factors in beige
- ZIP code individuals are color coded according to borough.


a) Variables' projection in PC1-2 (all)
(MCA on April 2018 NYC data)
Dim2 (7.2%)
Dim1 (18%)
Col.
Crime
SRCs
UrbInf_L
Sani_L
Traffic_L
IAO_L
Misdemeanor_OC
Felony_VH
HousCond_[298,1278]
NoiseBiz_H
NoiseResid_VH
Violation_H
NoiseTraf_H
NoiseTraf_M
NoiseConst_H
SocServ_MH
ConsumProt_H
HousCond_[137,298]
EnvProt_M
Violation_L
IAO_H
ConsumProt_M
Traffic_H
Felony_ML
NoiseResid_L
Felony_H
NoiseConst_M
NoiseBiz_VL
UrbInf_H
SocServ_VL
Traffic_VH
Misdemeanor_M
Sani_VH
WaterSyst_MH
NoiseConst_VL
ConsumProt_VL
WaterSyst_H
EnvProt_H
EnvProt_MH
WaterSyst_M
HousCond_[0,56]
NoiseTraf_VL
Violation_MH
IAO_MH
Traffic_M
Sani_M
NoiseResid_H
ConsumProt_L
IAO_M
NoiseTraf_L
NoiseBiz_L
NoiseConst_L
HousCond_[56,137]
Misdemeanor_H
UrbInf_H
Felony_M
SocServ_L
Sani_H
Violation_M
NoiseResid_M
WaterSyst_L
EnvProt_L
UrbInf_M
Violation_L
NoiseResid_L
Felony_ML
NoiseBiz_VL
SocServ_VL
NoiseConst_VL
ConsumProt_VL
HousCond_[0,56]
NoiseTraf_VL
Violation_M
NoiseResid_M


b) Variables' projection in PC1-2 (cos2>0.2)
(MCA on April 2018 NYC data)
Dim2 (7.2%)
Dim1 (18%)
Col.
Crime
SRCs
UrbInf_L
Sani_L
Traffic_L
IAO_L
Misdemeanor_OC
Felony_VH
HousCond_[298,1278]
NoiseBiz_H
NoiseResid_VH
Violation_H
NoiseTraf_M
SocServ_MH
IAO_H
Traffic_VH
Sani_VH
WaterSyst_L
EnvProt_L
Violation_L
NoiseResid_L
Felony_ML
NoiseBiz_VL
SocServ_VL
NoiseConst_VL
ConsumProt_VL
HousCond_[0,56]
NoiseTraf_VL
Violation_M
NoiseResid_M


a) Individuals' projection in PC1-2
(MCA on April 2018 NYC data)
Dim2 (7.2%)
Dim1 (18%)
cos2
0.75
0.50
0.25

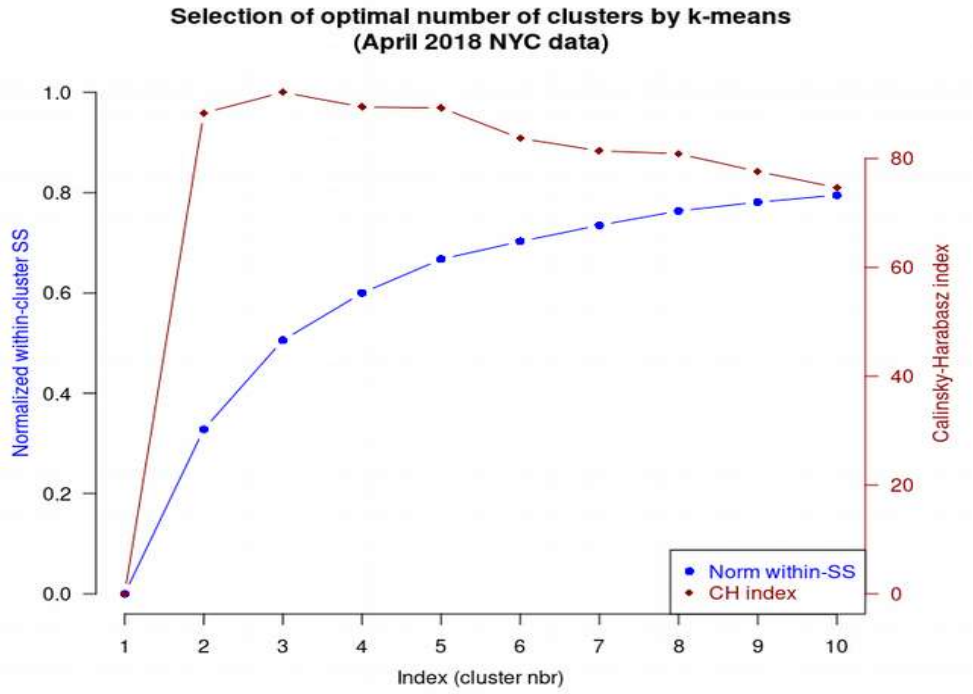

b) Individuals' projection in PC1-2, by NYC borough
(MCA on April 2018 NYC data)
Dim2 (7.2%)
Dim1 (18%)
Groups
Bronx
Brooklyn
Manhattan
Queens
Staten Island

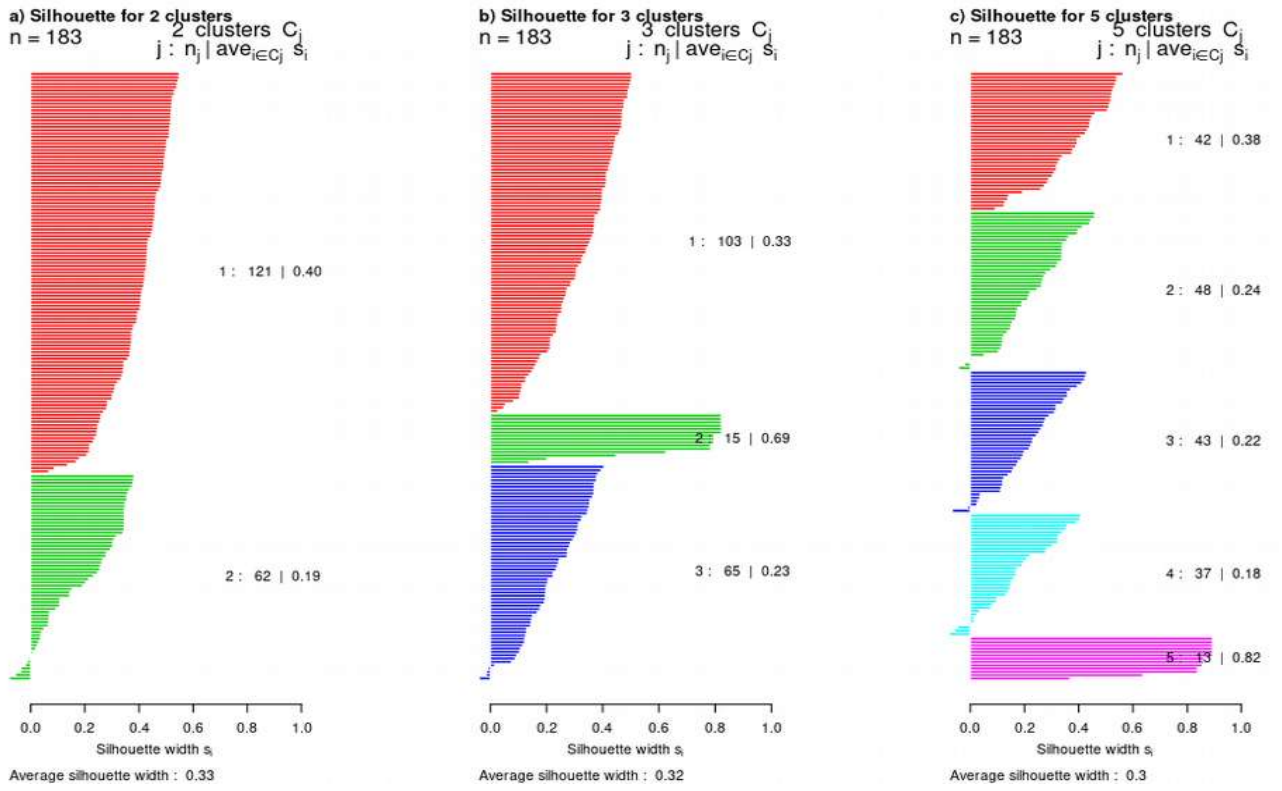


Identify ZIP codes in 2nd quadrant of PC12 var projection from MCA:

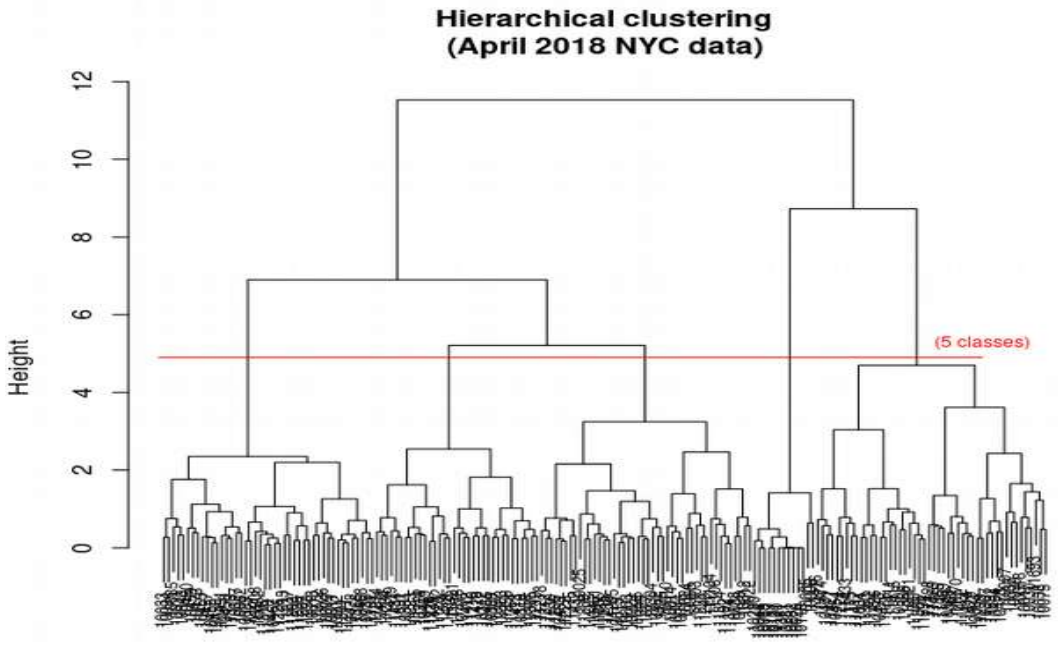
```
sum of all violation counts in 2nd quadrant: 3312
sum of all other violation counts: 2180

sum of all misdemeanor counts in 2nd quadrant: 12680
sum of all other misdemeanor counts: 7550

sum of all felonie counts in 2nd quadrant: 6430
sum of all other felony counts: 10272
```

Clustering analysis:

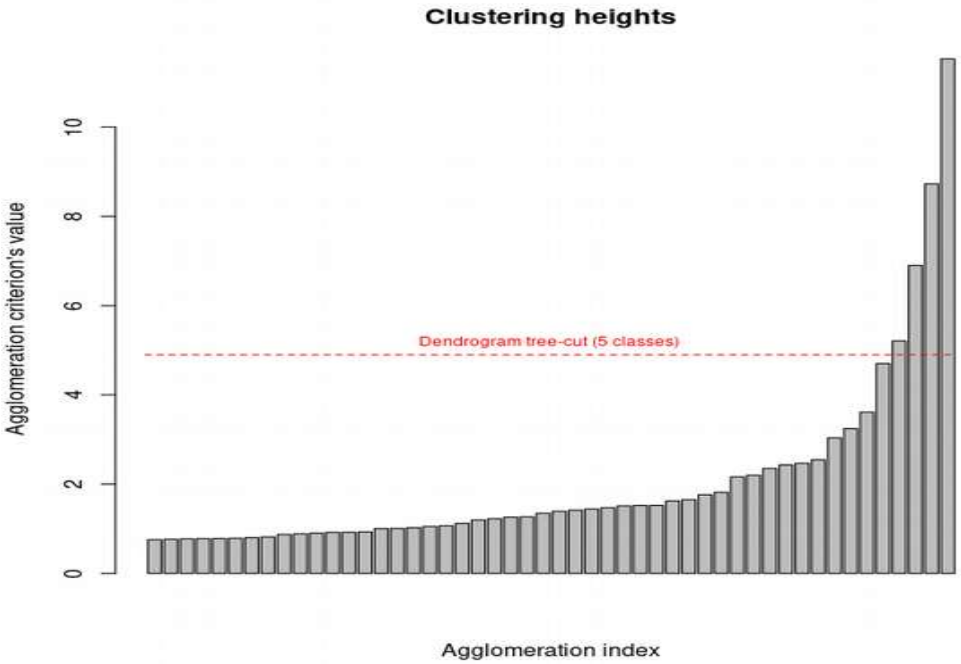


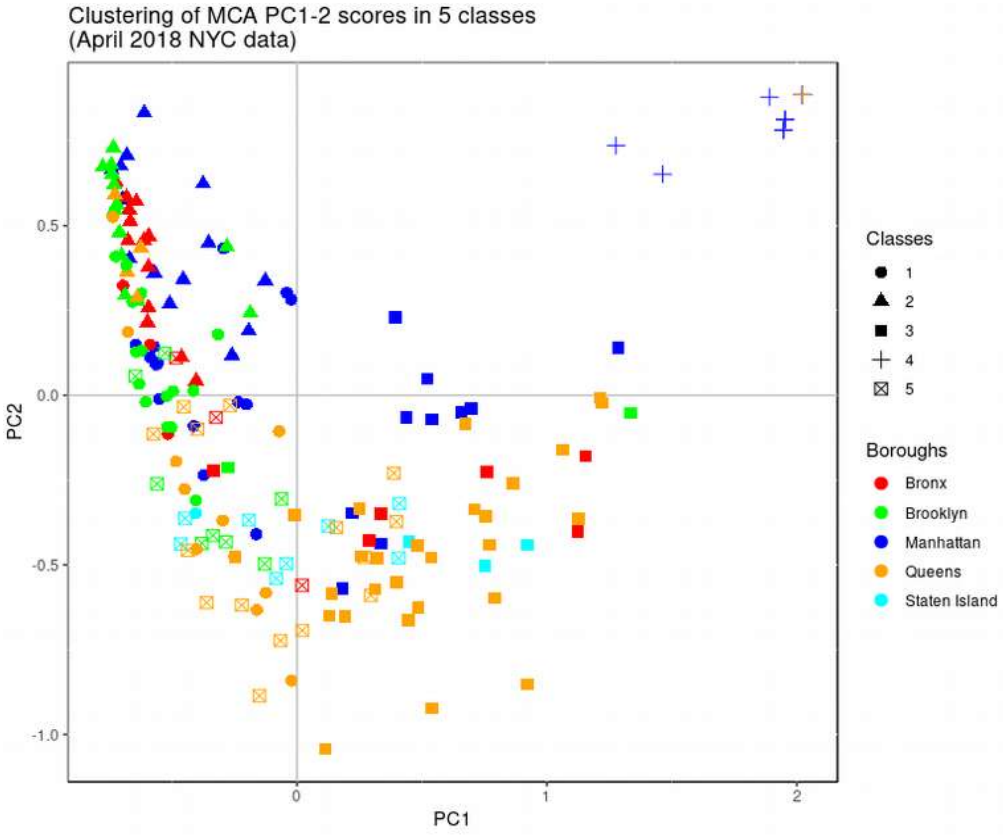


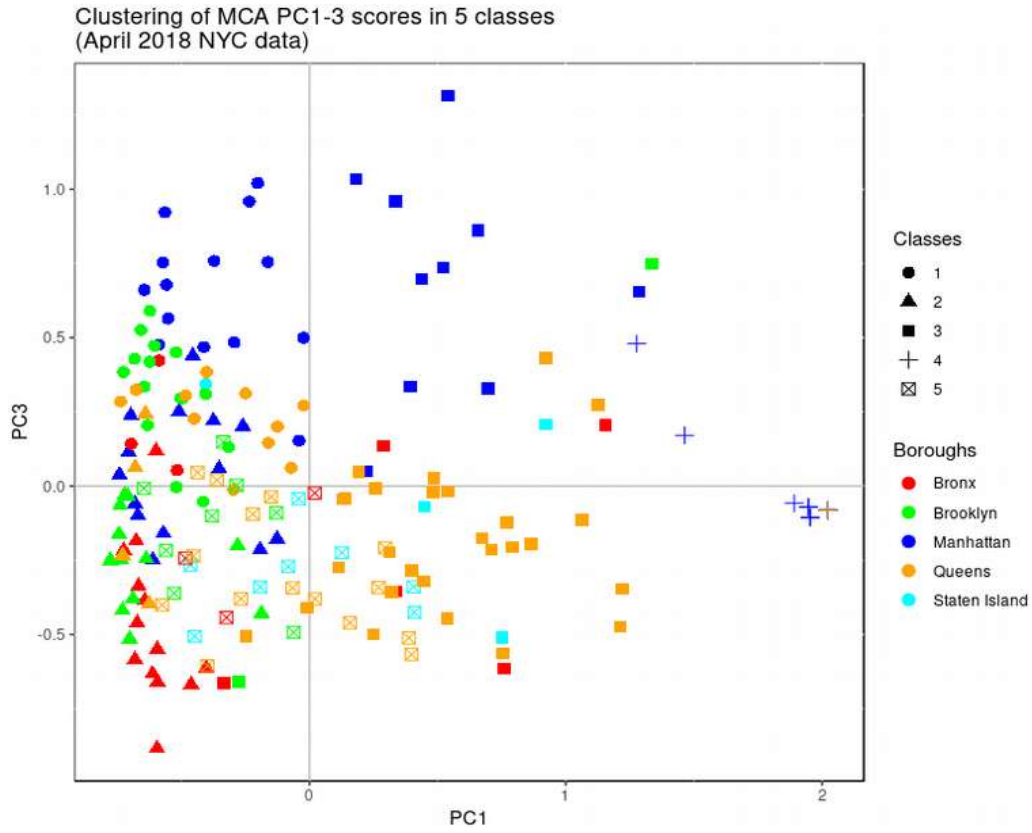
Choose 5 classes.

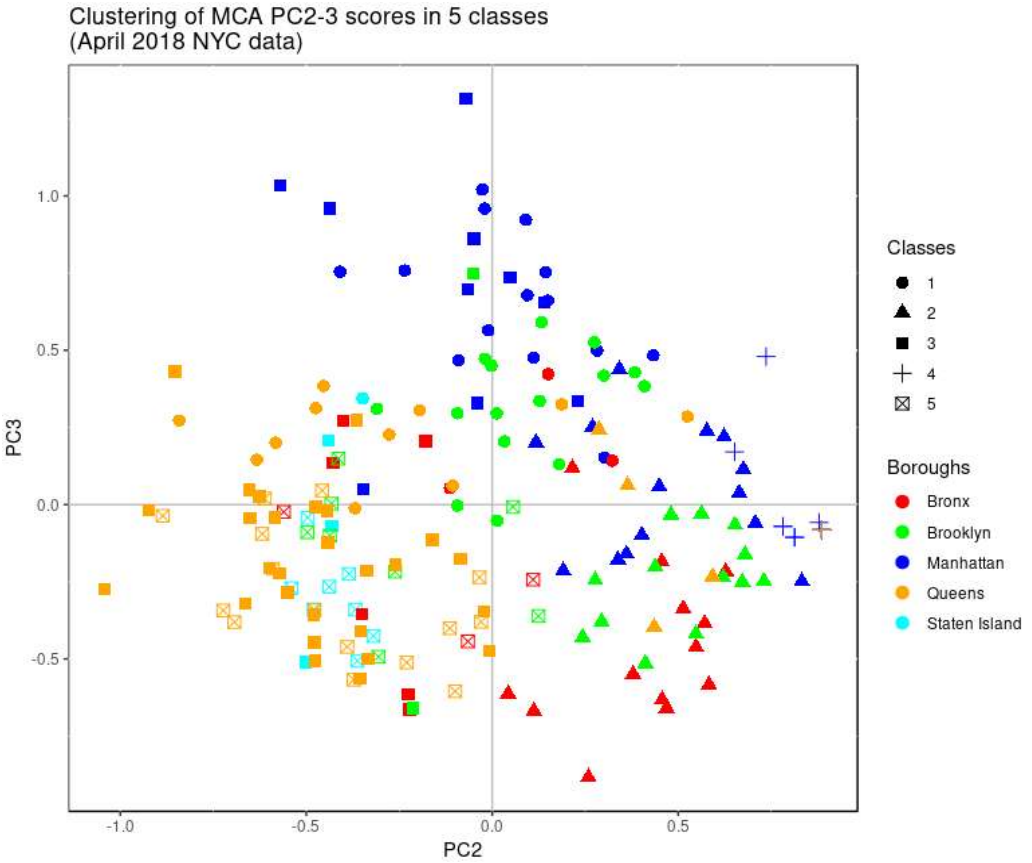


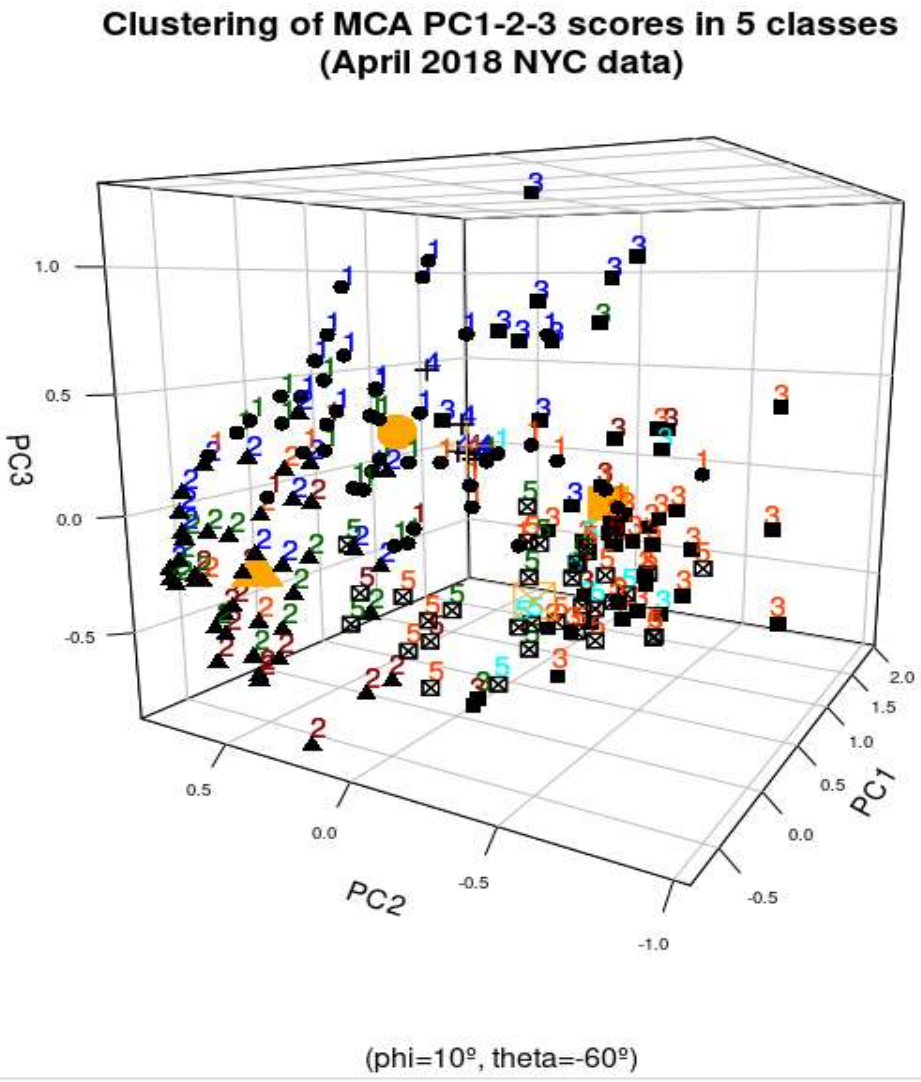
Clustering quality index for 5 classes, I_b=63.73


Clustering heights
Agglomeration criterion's value
0
2
4
6
8
10
Dendrogram tree-cut (5 classes)
Agglomeration index


Clustering of MCA PC1-2 scores in 5 classes
(April 2018 NYC data)
PC2
0.5
0.0
-0.5
-1.0
0
1
2
PC1
Classes
1
2
3
4
5
Boroughs
Bronx
Brooklyn
Manhattan
Queens
Staten Island


Clustering of MCA PC1-3 scores in 5 classes
(April 2018 NYC data)
PC3
PC1
1.0
0.5
0.0
-0.5
0
1
2
Classes
1
2
3
4
5
Boroughs
Bronx
Brooklyn
Manhattan
Queens
Staten Island


Clustering of MCA PC2-3 scores in 5 classes
(April 2018 NYC data)
PC3
PC2
1.0
0.5
0.0
-0.5
-1.0
-0.5
0.0
0.5
Classes
1
2
3
4
5
Boroughs
Bronx
Brooklyn
Manhattan
Queens
Staten Island



Contributions of each borough to inertia over 'nd' = 5 dimensions:

```
Bronx’ % (normalized) contribution to inertia:  11.6 %
Brooklyn's % (normalized) contribution to inertia:  15.4 %
Manhattan's % (normalized) contribution to inertia:  41.6 %
Queens' % (normalized) contribution to inertia:  26.4 %
Staten Island's % (normalized) contribution to inertia:  5 %
```

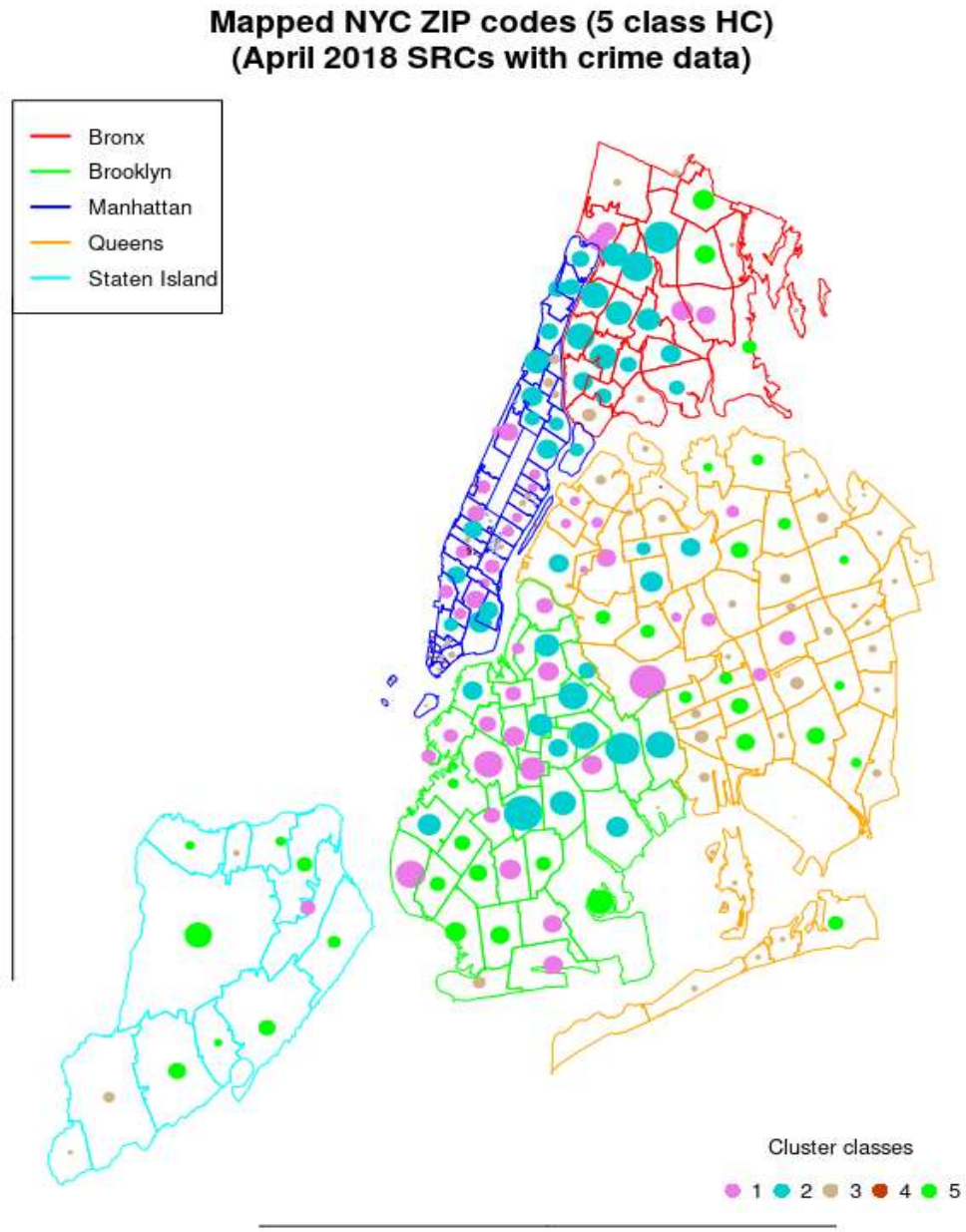
For more details (over 5 dim, non normalized):

```
Borough:  Bronx
 Number of ZIP codes:  25
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 4.6

Borough:  Brooklyn
 Number of ZIP codes:  38
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 6.1

Borough:  Manhattan
 Number of ZIP codes:  46
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 16.5
```

```
Borough:  Queens
 Number of ZIP codes:  59
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 10.5

Borough:  Staten Island
 Number of ZIP codes:  12
Borough's ZIPs' non-normalized % contribution to inertia over (5_dim): 2.0
```

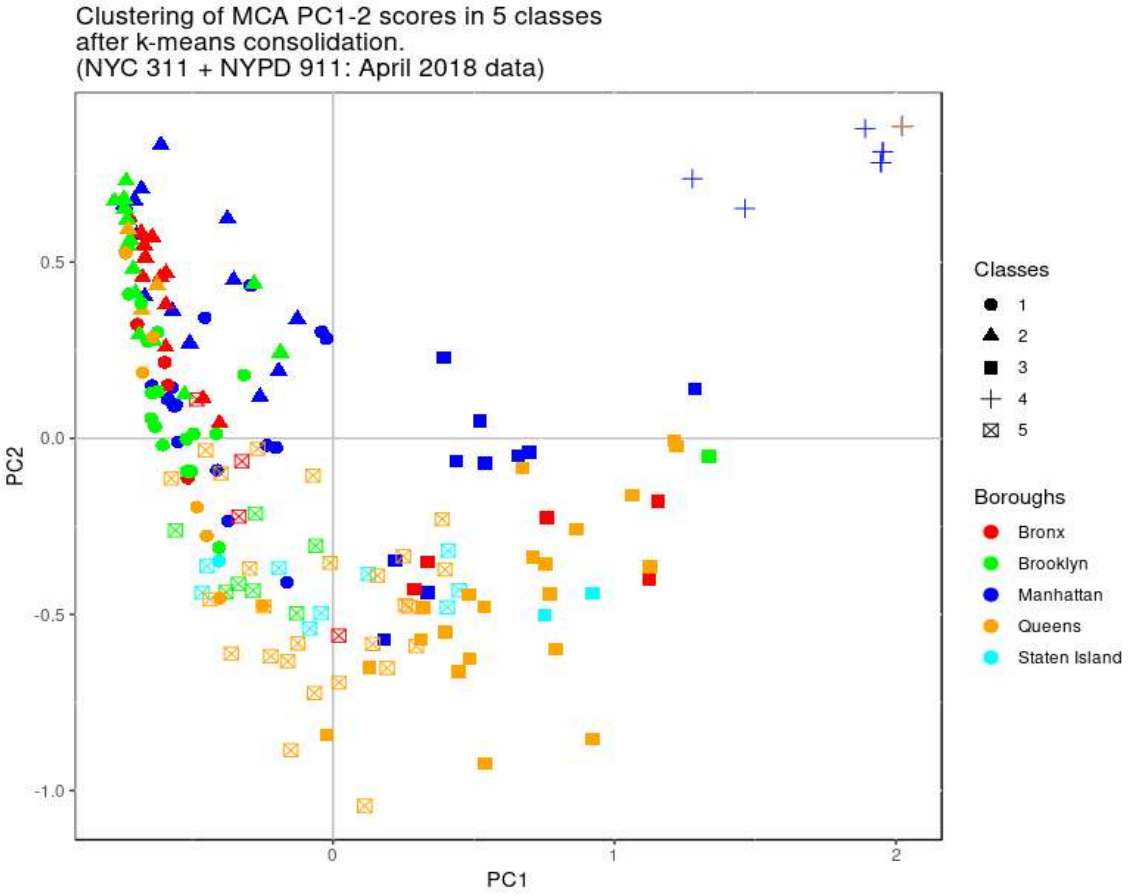
Contributions of each cluster class to inertia over 'nd' = 5 dimensions:

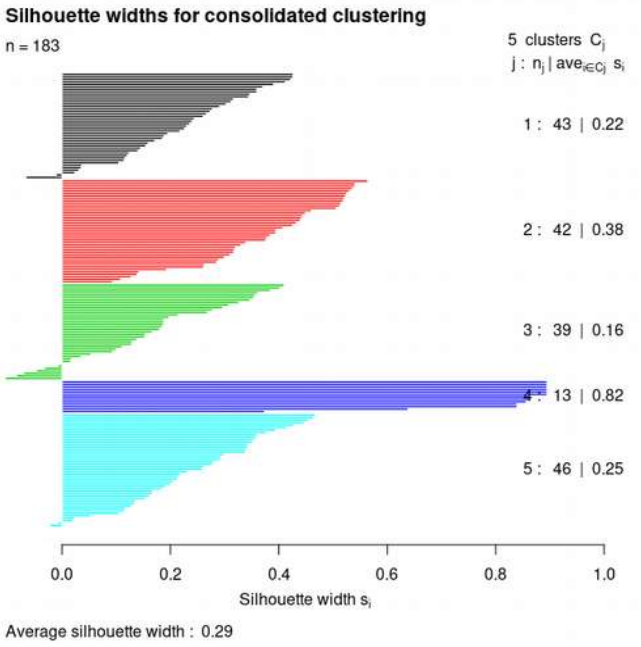
```
Cluster 1 's normalized contribution to inertia:  16.1 %
Cluster 2 's normalized contribution to inertia:  18.4 %
Cluster 3 's normalized contribution to inertia:  28.5 %
Cluster 4 's normalized contribution to inertia:  25.9 %
Cluster 5 's normalized contribution to inertia:  11.1 %
```



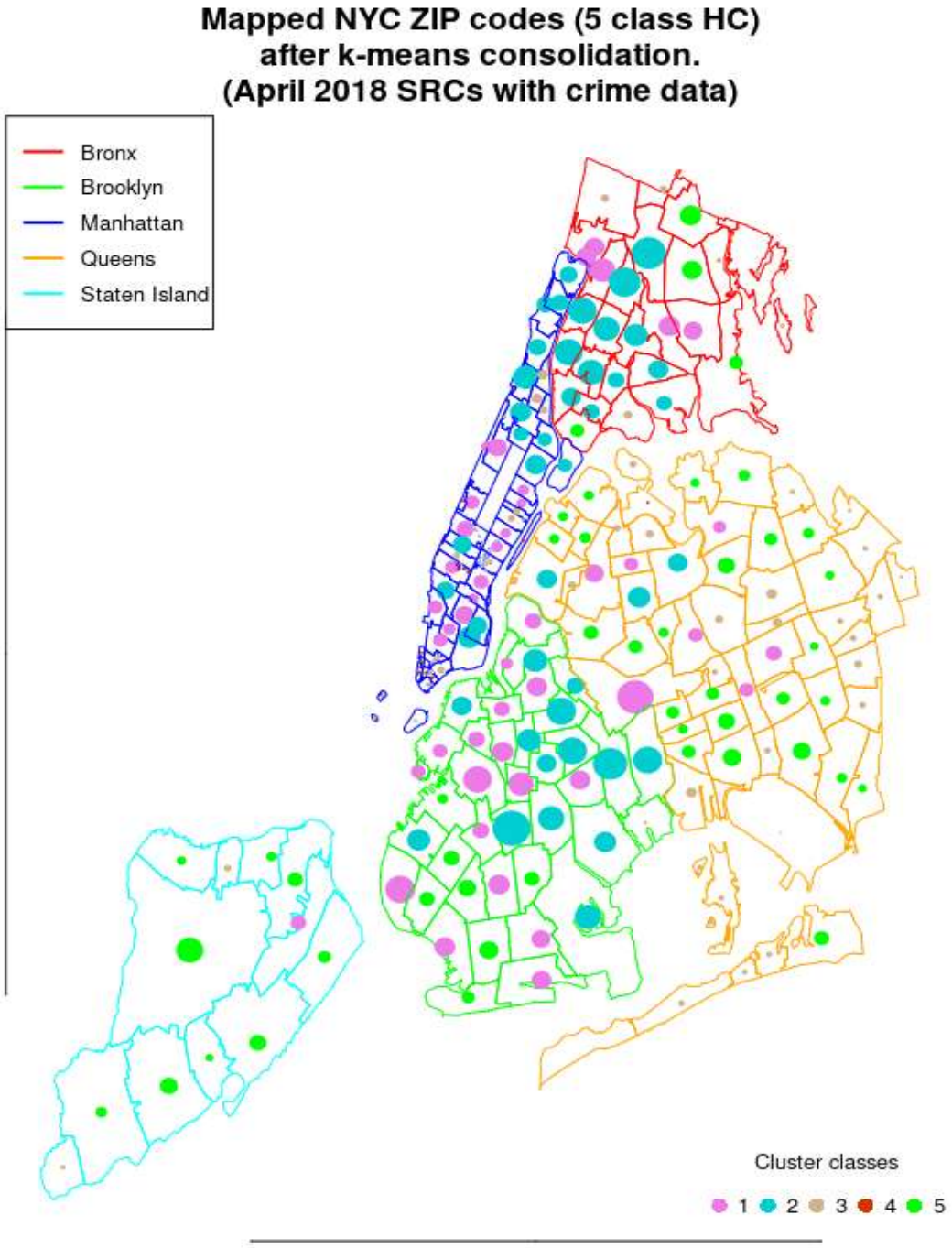
HC clustering with k-means consolidation:

Clustering quality index for 5 classes, after k-means consolidation, Ib = 66.75





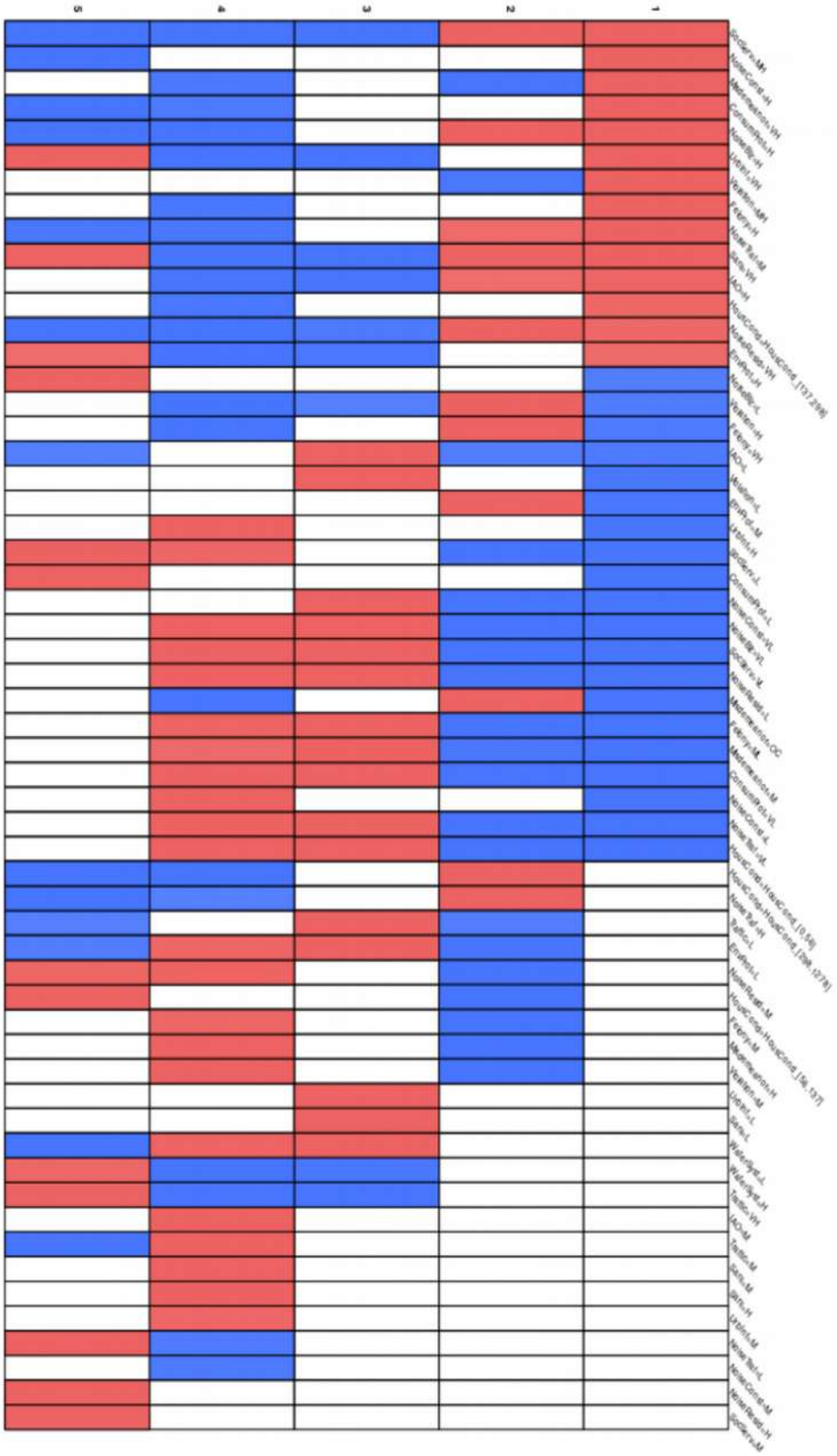

```
Cluster 1 's normalized contribution to
inertia:  15.9 %

Cluster 2 's normalized contribution to
inertia:  17.9 %

Cluster 3 's normalized contribution to
inertia:  24.9 %

Cluster 4 's normalized contribution to
inertia:  25.9 %

Cluster 5 's normalized contribution to
inertia:  15.4 %
```

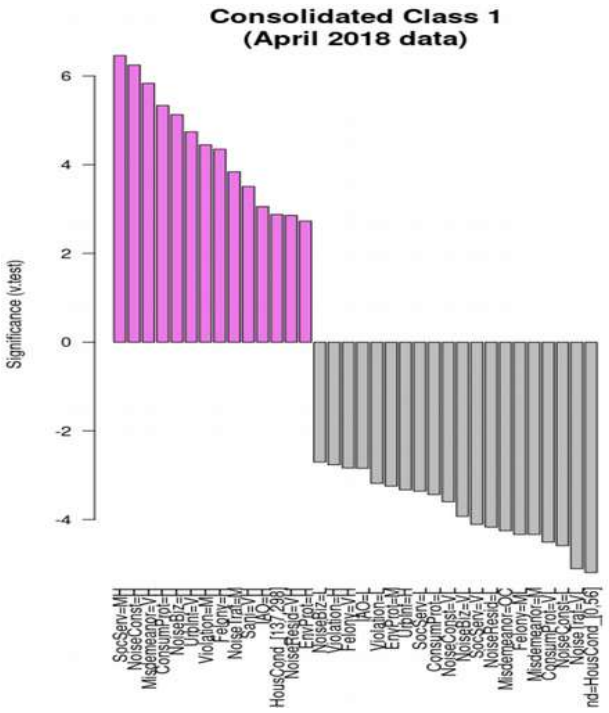


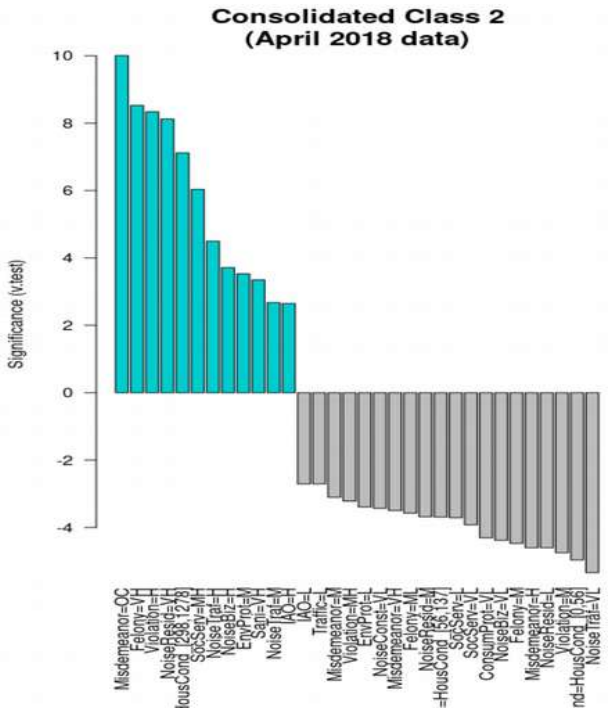
Var. **most** significantly related to construction of classes:
UrbInf Misdemeanor Sani NoiseResid HousCond

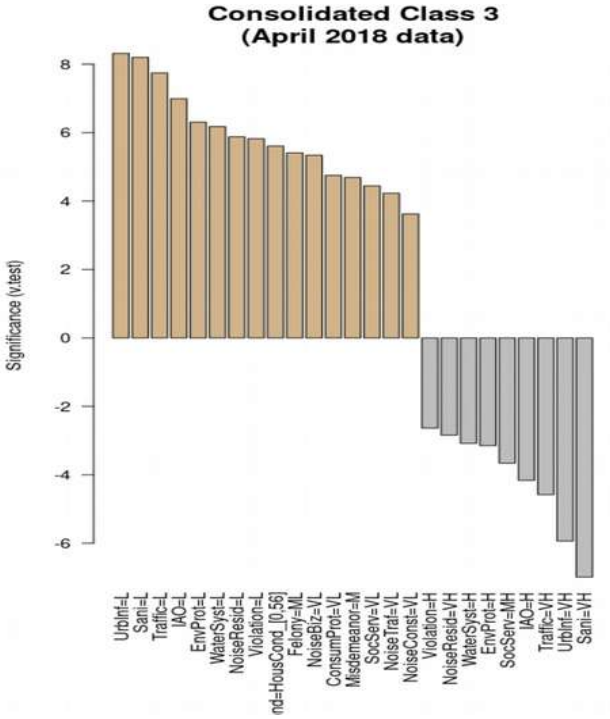
Var. **least** significantly related to construction of classes:
SocServ NoiseBiz NoiseConst EnvProt WaterSyst

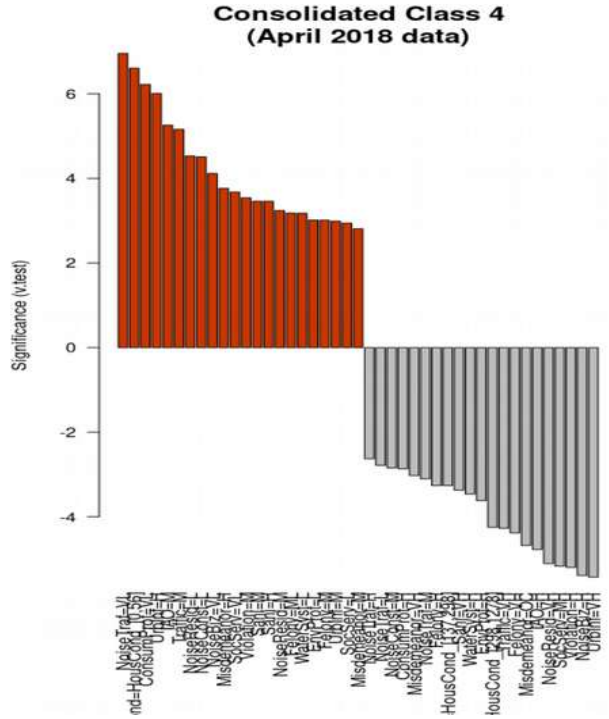
Clustering interpretation:

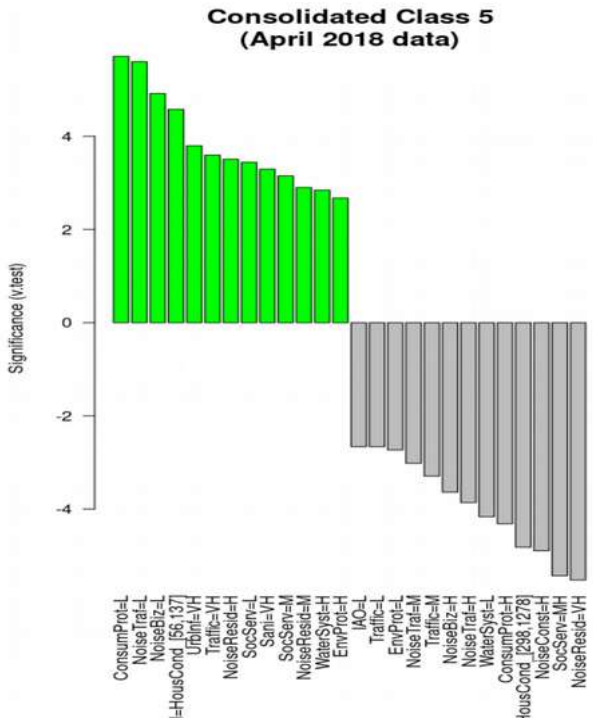



Consolidated Class 1
(April 2018 data)
Significance (v.test)
6
4
2
0
-2
-4
SocServ=MH
NoiseConst=H
Misdemeanor=VH
ConsumProt=H
NoiseBiz=H
UrbInt=VH
Violation=MH
Felony=H
NoiseTraf=M
Sani=VH
IAO=H
HousCond_[137,298)
NoiseResid=VH
EnvProt=H
NoiseBiz=L
Violation=H
Felony=VH
IAO=L
Violation=L
EnvProt=M
UrbInt=M
SocServ=L
ConsumProt=L
NoiseConst=VL
NoiseBiz=VL
SocServ=VL
NoiseResid=L
Misdemeanor=OC
Felony=M
Misdemeanor=M
ConsumProt=VL
NoiseTraf=VL
HousCond_[0,56]


Consolidated Class 2
(April 2018 data)
Significance (v.test)
10
8
6
4
2
0
-2
-4
Misdemeanor=OC
Felony=VH
Violation=H
NoiseResid=VH
HousCond_[298,1278]
SocServ=MH
NoiseTraf=H
NoiseBiz=H
EnvProt=M
Sani=VH
NoiseTraf=M
IAO=H
IAO=L
Traffic=L
Misdemeanor=M
Violation=MH
EnvProt=L
NoiseConst=VL
Misdemeanor=VH
Felony=ML
NoiseResid=M
HousCond_[56,137)
SocServ=L
SocServ=VL
ConsumProt=VL
NoiseBiz=VL
Felony=M
Misdemeanor=H
NoiseResid=L
Violation=M
HousCond_[0,56]
NoiseTraf=VL


Consolidated Class 3
(April 2018 data)
Significance (v.test)
8
6
4
2
0
-2
-4
-6
UrbInt=L
Sani=L
Traffic=L
IAO=L
EnvProt=L
WaterSyst=L
NoiseResid=L
Violation=L
HousCond_[0,56]
Felony=ML
NoiseBiz=VL
ConsumProt=VL
Misdemeanor=M
SocServ=VL
NoiseTraf=VL
NoiseConst=VL
Violation=H
NoiseResid=VH
WaterSyst=H
EnvProt=H
SocServ=MH
IAO=H
Traffic=VH
UrbInt=VH
Sani=VH


Consolidated Class 4
(April 2018 data)
Significance (v.test)
6
4
2
0
-2
-4



Class 1 's most significant modalities:
SocServ=MH NoiseConst=H Misdemeanor=VH ConsumProt=H NoiseBiz=H UrbInf=VH Violation=MH Felony=H NoiseTraf=M Sani=VH IAO=H HousCond=[137,298] NoiseResid=VH EnvProt=H

Class 2 's most significant modalities:
Misdemeanor=OC Felony=VH Violation=H NoiseResid=VH HousCond=[298,1278] SocServ=MH NoiseTraf=H NoiseBiz=H EnvProt=M Sani=VH NoiseTraf=M IAO=H

Class 3 's most significant modalities:
UrbInf=L Sani=L Traffic=L IAO=L EnvProt=L WaterSyst=L NoiseResid=L Violation=L HousCond=[0,56] Felony=ML NoiseBiz=VL ConsumProt=VL Misdemeanor=M SocServ=VL NoiseTraf=VL NoiseConst=VL

Class 4 's most significant modalities:
NoiseTraf=VL HousCond=[0,56] ConsumProt=VL UrbInf=H IAO=M Traffic=M NoiseResid=L NoiseConst=L NoiseBiz=VL Misdemeanor=H SocServ=VL Violation=M Sani=M Sani=H NoiseResid=M Felony=ML WaterSyst=L EnvProt=L Felony=M UrbInf=M SocServ=L Misdemeanor=M

Class 5 's most significant modalities:
ConsumProt=L NoiseTraf=L NoiseBiz=L HousCond=[56,137] UrbInf=VH Traffic=VH NoiseResid=H SocServ=L Sani=VH SocServ=M NoiseResid=M WaterSyst=H EnvProt=H